THIS WEEK

EDITORIALS

LIBEL The UK parliament signs off on long-awaited legal reforms **p.6**

WORLD VIEW The dismal sexism at the top of Australia's science **p.7**

FEATHERS Pretty plumage is painted by cell-signalling numbers **p.9**

Plan for the future

The White House urgently needs to set out a clear plan for how it intends to monitor the state of Earth.

President Barack Obama's administration released its first national strategy for civil Earth observations on 19 April. The report comes six years after the US National Research Council (NRC) warned that inadequate funding and mismanagement had put 'at great risk' the United States' ability to monitor Earth from space. The strategy does little to reassure.

The 60-page document, written by a federal task force, lays out a process to determine the types of observations that deserve government support. But it does not provide what is most urgently needed: clear and specific guidance from the White House on what the government considers to be the most important Earth-science satellite missions — or when they should be launched.

That type of plan, long overdue, grows more important as the fiscal crisis deepens and the demand for such observations rises (see page 13). Meanwhile, the country's ageing collection of Earth-observing satellites continues its long decline. The number of US probes is likely to dwindle from 23 to just 6 by 2020, threatening to degrade scientists' ability to track climate change, forecast weather and monitor natural disasters.

Obama is one of many to blame for the brewing crisis. The lack of leadership at the White House is matched by the intransigence of Congress, which set in motion the across-the-board sequestration spending cuts that took effect on 1 March, slashing about 5% from the budgets of NASA, the National Oceanic and Atmospheric Administration and other key science agencies.

Lawmakers also approved steep cuts to NASA's Earth-science budget beginning in 2002, as then-President George W. Bush foolishly directed the space agency to focus instead on manned missions to the Moon and Mars. Obama has pushed Congress to reverse that decline, but the programme's budget — US\$1.8 billion this year — still falls well below the \$2 billion-per-year target that the NRC says is necessary to launch 17 'high-priority' missions by 2020. That makes a blunt discussion about what level of future spending is reasonable and achievable all the more urgent.

The situation is little better in Europe. Member states approved a budget last year that gives the European Space Agency about 80% of what it is seeking to develop research satellites over the next five years. Scientists are worried that the shortfall could delay the planned launch of a climate-change mission, called Earth Explorer 8, in 2018.

The US government has also been forced to cope with plain bad luck. The Orbiting Carbon Observatory, a much-anticipated satellite designed to track the level of carbon dioxide in the atmosphere, crashed shortly after launch in 2009. Two years later, a similar failure claimed Glory, a mission to monitor Earth's energy balance, before it could reach orbit. The two incidents cost NASA more than \$700 million — not including the \$470 million or so the space agency is spending to launch a copy of the observatory in 2014.

To its credit, the Obama administration has made some progress to improve the nation's eyes in the sky. NASA successfully launched the polar-orbiting climate and weather satellite Suomi National Polar-orbiting Partnership in October 2011 and the ocean-salinity mission Aquarius in June of that year. Last month, Landsat-8 reached orbit, ensuring that the world's longest-running global-change data set will continue.

Yet the long-term forecast for US Earth observations remains grim. The US government plans to launch just six satellites between 2014 and 2020, including only two of the four missions that the NRC panels deemed the most important. The other two — designed to measure long-term changes in solar radiation, ice-sheet velocities and terrestrial biomass — have been shelved indefinitely by the White House.

"The long-term forecast for US Earth observations remains grim."

Researchers who warned for years of this slow-moving disaster are now left to watch it unfold. And it comes at a time when concern is growing about the pace of climate change and the pressure that the world's burgeoning population is placing on limited natural resources.

Obama's science adviser, John Holdren, says that the administration will release a detailed national plan for Earth-observing missions as supplement to the White House budget request delivered to Congress on 10 April. It cannot come soon enough. Progress depends on the United States making hard decisions about what Earth observations it needs and how best to provide them. For scientists, and society, the dilemma is clear: we cannot manage what we cannot measure. ■

Fields of gold

Research on transgenic crops must be done outside industry if it is to fulfil its early promise.

It was 30 years ago this month that scientists first published the news that they could place functional foreign genes into plant cells. The feat promised to launch an exciting phase in biotechnology, in which desired traits and abilities could be coaxed into plants used for food, fibres and even fuel. Genetically modified (GM) crops promised to make life easier and nature's bounty even more desirable.

As a series of articles in this week's *Nature* explores, things have not worked out that way (see page 21). The future matters more than the past, but when it comes to GM crops, the past is instructive.

Soon after the 1983 breakthrough, biotechnology companies developing GM crops became hugely attractive to investors. Calgene in Davis, California, for example, developed the Flavr Savr tomato — engineered

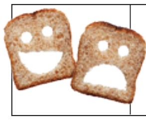
to remain firm after ripening — which captured attention, especially when the iconic Campbell Soup Company invested in its development. Like many at the time, Campbell's was fascinated by the promise that tomatoes could be ripened on the vine to accentuate their flavour and still make the trip to the supermarket and the dinner table without turning to mush.

In early 1992, analysts predicted regulatory approval for the GM tomato within a month, and a market of at least US$500 million a year. But less than a decade after their birth, GM crops were already facing a difficult adolescence. What was once deemed biological wizardry was increasingly being labelled Frankenfood. Consumers in Europe were bristling at the aggressive marketing of GM giant Monsanto, based in St Louis, Missouri. The Flavr Savr suffered more than a year of delays at the US Food and Drug Administration, and Campbell's began to state that it had no intention of putting the tomatoes in its soups without approval from the public. What had gone wrong? According to one analyst quoted at the time, the biotech sector had failed to prepare consumers appropriately: "Now, they realize that they have to be articulate and educate an uninformed public."

The Flavr Savr was approved in 1994 but never took off commercially. In the meantime, the biotech industry had shifted much of its attention to traits that aimed not to delight consumers, but rather to increase farm yields. Herbicide-tolerant and pest-resistant crops proliferated in the United States and more than two dozen other countries. GM organisms were to become agricultural tools.

In many places where they are planted, these GM crops have replaced conventional planting almost entirely. Yields and profits have increased, farmers have been generally happy to adopt the transgenic seeds and the technology has even made good on some of its promises to help the environment by reducing the amount and variety of pesticides needed.

GM crops, of course, still face a public-relations problem. Fears of the unfamiliar and 'unnatural', and concerns about health or environmental impacts, have frequently prevented approval and adoption of the crops, especially in Europe, where protesters have destroyed experiments. The United States, the world's most active user of GM crops, has seen renewed backlash as calls grow for foods with GM ingredients to be clearly labelled.



GM CROPS: PROMISE & REALITY
A *Nature* special issue
nature.com/gmcrops

The analyst who spoke of an uninformed public may have been correct in 1993, but such a claim no longer applies. People are positively swimming in information about GM technologies. Much of it is wrong — on both sides of the debate. But a lot of this incorrect information is sophisticated, backed by legitimate-sounding research and written with certitude. (With GM crops, a good gauge of a statement's fallacy is the conviction with which it is delivered.)

Armed with misinformation, debaters have taken to the streets, the supermarkets and social media. With a topic as sensitive and dear to people as the food they eat and give to their children, those who play to the fears, concerns and uncertainty surrounding GM crops often seem to have the upper hand. And the fears are entwined with mistrust of the seed companies. Supporting GM crops can seem a thankless job: it is worthwhile to stand up for good science and the promise that it holds, but defending profit-hungry corporations feels less rewarding.

Still, there is reason to stand up for the continued use and development of GM crops. Genetic modification is a nascent technology for which development has moved very quickly to commercialization. That has forced most research into the for-profit sector. Without broader research programmes outside the seed industry, developments will continue to be profit-driven, limiting the chance for many of the advances that were promised 30 years ago — such as feeding the planet's burgeoning population sustainably, reducing the environmental footprint of farming and delivering products that amaze and delight. Transgenic technologies are by no means the only way to achieve these aims, but the speed and precision that they offer over traditional breeding techniques made them indispensable 30 years ago. They still are today. ■

Freed speech

The reform of English libel law is a victory, even if it doesn't achieve everything that was hoped.

In a typically British piece of formal pomp, the speaker of the UK House of Commons, John Bercow, last week declared: "I have to acquaint the House that the House has been to the House of Peers, where a Commission under the Great Seal was read, authorizing the Royal Assent to the following Acts."

In the list of new legislation that followed, alongside the 'Marine Navigation (No. 2) Act' and the 'Groceries Code Adjudicator Act', Bercow announced the Queen's formal approval of a long-awaited reform to libel laws in England and Wales.

Nature was taken to court under the previous version of these laws, which were widely regarded as skewed in favour of those who claim libel, and we were among the many supporters of the Libel Reform Campaign, which drove the fight for change. Cases such as that of science writer Simon Singh, who was forced to defend himself against a claim by the British Chiropractic Association over an article published in the *Guardian* newspaper in 2008, galvanized the public and raised concern about the laws' chilling effects on the free expression of scientific opinion.

Those cases ended in victories for Singh, for *Nature* and for scientific debate and free speech. But it was rightly feared that those without the resources of *Nature* or the tenacity of Singh would back down rather than face the costs of going to court, or might even shy away from making statements that might attract attention from litigious parties in the first place.

The new law will require that bodies that trade for profit show "serious financial loss" if they wish to sue someone for defamation. It also includes formalized defences for journalists publishing on matters of public interest, and further protections for the reporting of statements made in peer-reviewed journals and at international conferences.

'Libel tourism' — in which those with no real link to Britain come to use the unfair laws in London courts — will be restricted by the new act. It sets bars for action against people who do not live in the United Kingdom or the rest of Europe, unless the claimant can show that England is truly the most appropriate venue for legal action.

These are all real gains that should improve the communication of science by making it easier to speak truths that some may not wish to hear.

The rewriting of the law led to celebration among the scientists, journalists, lawyers and others who have pushed for reform. But there were cautionary voices. It is not yet clear how the new law will work in practice for much of the Internet. And it may not reduce the cost of litigation. If defending an action is still financially crippling, concerns that the law can be used to threaten people into silence will persist.

Robert Dougans, solicitor-advocate at the litigation firm Bryan Cave in London, who represented Simon Singh in his fight with the British Chiropractic Association, said, "Frankly, I cannot see this having made any difference in any case I have been involved in, and I wish an opportunity had been taken to re-think defamation law *ab initio*." (See *Nature* http://doi.org/mc6; 2013.)

Dougans may be too pessimistic. There is good reason for those who have fought hard to rejoice. But it remains to be tested whether the culture of suppression has truly been swept away. If it has not, the fight will have to begin again. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

WALTER AND ELIZA HALL INST.

Australian science needs more female fellows

The Australian Academy of Science must take urgent steps to address the lack of gender equality among its elected fellows, warns **Douglas Hilton**.

The Australian Academy of Science (AAS) is a hall of fame for Australian scientists. To be elected as a fellow of the taxpayer-funded academy, which was modelled on the Royal Society in the United Kingdom, is a high honour indeed. So why does the AAS treat female scientists with such disdain?

Of the almost 500 living fellows of the academy, 92% are men and 8% are women. Given that fellows have been elected over the past 60 or so years, and science has historically been a male-dominated profession, this imbalance is not altogether surprising. Indeed, the problem of gender inequality in science is an international issue (see go.nature.com/zzexkh). With the increasing participation of women in science since the 1970s and with some fields containing a majority of women as undergraduate and postgraduate students for many years, one would expect the situation to be changing for the better every year — but this is not the case in Australia.

In fact, 2013 represented a low point in the history of the AAS: not one of the 37 candidates shortlisted for election was a woman, and so none of the 20 newly elected fellows was a woman. To put it another way, the academy believed that there were at least 37 male candidates more worthy than the best female candidate. This is disappointing enough, but perhaps the greater scandal is that there was no acknowledgement that this was even a problem. Unfortunately, 2013 was not exceptional: in 6 of the past 12 years, only one female member has been elected and, as a consequence, improvement in the overall gender balance has been slow.

WHY DOES THE AAS TREAT **FEMALE** SCIENTISTS WITH SUCH **DISDAIN?**

It is farcical that in 2013 the academy could not find a single woman whom it deemed worthy of election. Clearly, the processes and procedures that it uses to find, consider and elect fellows are so flawed that a complete overhaul is required.

The academy does go to a great deal of trouble to ensure equal participation of researchers from the physical and biological sciences. Its 13 membership committees are structured around scientific disciplines and ensure that, almost without exception, between one and three new members from each area are elected each year. If the top five candidates for election come from a single discipline, then tough luck — two will miss out to less-competitive candidates from other disciplines. The academy rightly believes that inclusive participation of researchers across fields is essential to its long-term vibrancy and viability, and that quotas are an acceptable means of achieving this outcome. What a shame that its inclusive policies do not stretch to gender.

A key recommendation from the 2005 Australian government review of the Australian learned academies was that academies should "focus on addressing gender imbalances in their fellowships". This should have been a call to action, a spur to decisive and creative policy changes; however, this was not the case. The government has in effect placed the academy on notice, with its funding now at risk.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/dbtfck

There are several ways to remedy the situation.

Ironically, the academy recently released a blueprint to tackle the gender-equality problem in science — *Gender Equity: Current Issues, Best Practice and New Ideas*. It recommended that a university or institute should receive government funding only if it provides evidence that it has a functional gender-equity committee. This is a proposal that the academy itself should adopt. It should create a standing committee, with the chair of that committee taking a seat on its governing council. This committee would ensure that as broad a range of talented women as possible are nominated for election to the academy and to improve gender-equity procedures.

A second change is a variant of the quota system that the academy has long employed to ensure disciplinary diversity. One option would be to limit the number of new male members to the number of female members elected. In this way, there would be equal numbers of men and women elected every year, and the overall gender balance would improve over time. Despite the fact that the academy has been willing to use similar tools to ensure discipline diversity for more than 60 years, to promote female scientists in this way would no doubt generate howls of protest. Critics say that it will produce a two-tier system, in which women will be viewed, and might view themselves, as 'second-class citizens'. In reality, however, only outstanding women will be elected, and perhaps the only cost will be that fewer men will be elected each year.

A third innovation that the academy must embrace is to be open about its problem. One of the most disappointing aspects of this year's election was that among the press releases and fanfare about election of new members and the wonderful state of Australian science, there was no public acknowledgement that the failure to find a single woman worthy of election is even a problem.

Addressing this issue is not rocket science. Surely an organization that includes the best mathematicians, chemists, physicists and biologists — and, yes, rocket scientists — can find the time, energy, imagination, passion and intellect to bring the hall of fame of Australian science into the modern age. ■

Douglas Hilton *is director of The Walter and Eliza Hall Institute of Medical Research and head of the Department of Medical Biology at the University of Melbourne, Australia, and a fellow of the Australian Academy of Science and the Australian Academy of Technological Sciences and Engineering.*
e-mail: hilton@wehi.edu.au

RESEARCH HIGHLIGHTS

Selections from the scientific literature

URBAN GROWTH

The shape of cities to come

The physical features of the world's largest cities are shifting as urban centres in Asia expand upwards and outwards.

Steve Frolking at the University of New Hampshire in Durham and his colleagues combined satellite observations of night-time illumination and urban backscatter — the reflection of microwaves from the surface of built-up land — to infer how the structure of 100 large cities around the world changed from 1999 to 2009. Most Chinese cities grew vertically, echoing drastic increases in land prices. However, cities in India and Africa expanded outwards, owing to factors such as unplanned urban development and building-height limits.

The data could be used to help researchers to understand how urbanization affects energy consumption and greenhouse-gas emissions, the authors note.

Environ. Res. Lett. **8, 024004 (2013)**

GENE THERAPY

Enzymes fix disease genes

Genome-editing enzymes can be used to correct patient-specific genetic defects.

Mutations in the *COL7A1* gene cause the skin-blistering disease recessive dystrophic epidermolysis bullosa, which can be fatal. A team led by Jakub Tolar at the University of Minnesota in Minneapolis engineered enzymes called transcription activator-like effector nucleases (TALENs) to repair the defective part of the gene in skin cells from a patient with the disease.

These corrected cells were then reprogrammed into pluripotent stem cells, which, the authors suggest, could be developed into more-specialized cells that are suitable for therapy.

Mol. Ther. **http://dx.doi.org/10.1038/mt.2013.56 (2013)**

ELISABETTA VISALBERGHI

ARCHAEOLOGY

Monkeys make their mark

Wild monkeys leave behind long-lasting traces of their efforts to crack nuts with tools — evidence that could be useful to archaeologists.

Bearded capuchin monkeys (*Sapajus libidinosus,* pictured) open seeds such as palm nuts by placing them on stone and wood surfaces, or 'anvils', and then pounding them with rocks. Elisabetta Visalberghi at the Institute of Cognitive Sciences and Technologies in Rome and her colleagues tracked this behaviour every month for three years at a study site in Brazil. The monkeys left behind tell-tale pits on the surfaces, and the discarded nut shells remained strewn around the anvils for years. The animals also moved stone hammers to use them at other anvils. The physical records left by modern monkeys create archaeological signatures that could help researchers to study tool use among ancient monkeys and early humans.

J. Archaeol. Sci. **http://dx.doi.org/10.1016/j.jas.2013.03.021 (2013)**

DISEASE RESEARCH

New hormone for diabetes

A newly identified hormone stimulates growth of insulin-producing cells in the mouse pancreas.

Douglas Melton and his colleagues at Harvard University in Cambridge, Massachusetts, identified the hormone, which they call betatrophin, by searching for genes that became more active in fat and liver tissue when insulin signalling was blocked. Injection of other mice with betatrophin resulted in an average 17-fold increase in rates of proliferation for β cells, the cells that deteriorate in some forms of diabetes. Although the hormone's mechanism of action is not yet known, the researchers did show that human livers also produce betatrophin. The hormone might one day replace insulin as a treatment for diabetes, the authors say.

Cell **http://dx.doi.org/10.1016/j.cell.2013.04.008 (2013)**

For a longer story on this research, see go.nature.com/5esyqp

CONSERVATION BIOLOGY

Bigger is better for protecting seas

Although small protected marine zones are important for conservation, they do not protect fish as well as larger areas do.

Nicholas Graham at James Cook University in Townsville, Australia, and Tim McClanahan at the Wildlife Conservation Society in New York assessed biomass and composition of coral-reef fish in several marine protected areas in the Indian Ocean. The authors compared small no-take zones — areas less than 10 square kilometres that were protected from human disturbance — with much larger marine 'wildernesses'. At a depth of 9 metres, mean fish biomass in the 640,000-square-kilometre Chagos Archipelago weighed in at six times the amount per hectare than that of the most successful small

no-take zone. Coral reef wildernesses in the Caribbean Sea and Pacific Ocean were similarly heavy with biomass.
Bioscience **http://dx.doi.org/10.1525/bio.2013.63.5.13 (2013)**

PALAEONTOLOGY

Winged raptor dined on fish

A fossil of a dinosaur that was thought to feed on tree-living animals has been found with a fish in its belly.

Used to understand the origins of flight, fossils of the four-winged feathered raptor *Microraptor gui,* which lived 120 million years ago have previously been found with a bird and a potentially tree-climbing mammal preserved in their guts.

Scott Persons at the University of Alberta in Edmonton, Canada, and his colleagues report on a fossil (**pictured**) containing a partially digested fish. The authors also describe adaptations — such as front teeth that project forward — that are similar to those seen in fish-hunting animals.

The feeding habits of *Microraptor* spp are now the best sampled of any non-avian dinosaur, revealing it as a generalist predator in arboreal and aquatic habitats, the authors say.
Evolution **http://dx.doi.org/10.1111/evo.12119 (2013)**

L. FANG

QUANTUM MECHANICS

Exchange-free communication

Researchers have proposed a mode of quantum communication whereby information could travel between two parties without the exchange of physical particles.

Quantum communication holds the promise of ultrasecure encryption, but most schemes proposed so far require the communicating parties to exchange particles. Hatim Salih at King Abdulaziz City for Science and Technology in Riyadh, Saudi Arabia, and his colleagues suggest a scheme in which a photon kept by one party is influenced by the other party opening or closing a channel between them. In principle, this allows for a measurement that can securely convey information without transferring or exchanging physical particles. This challenges long-standing assumptions of the requirements for communication, the authors say.
Phys. Rev. Lett. 110, **170502 (2013)**

NEUROSCIENCE

Secrets of brain building revealed

Separate studies reveal the mechanisms by which brain cells assume their rightful places.

Magdalena Götz at the University of Munich in Germany and her colleagues manipulated a single DNA-associated protein to promote folding in a normally smooth region of mouse brain. Low levels of the protein cause cells to divide along one plane, and high levels cause division along another. Regulation of the protein, in turn, permits complex folding that creates more room for the cerebral cortex, the layer of neural tissue covering the cerebrum.

Meanwhile, researchers led by Dwight Bergles at Johns Hopkins University School of Medicine in Baltimore, Maryland, showed how precursor cells maintain a constant density of neural support cells as these cells differentiate and die. The team used time-lapse imaging of adult mouse brains to reveal that the mobile precursor cells constantly survey their environments and avoid each other, establishing a grid-like distribution throughout the nervous system.
Cell 153, **535–549; *Nature Neur.* http://dx.doi.org/10.1038/nn3390 (2013)**

COMMUNITY CHOICE

The most viewed papers in science

GENOMICS

Reading tangled RNA sequences

HIGHLY READ on genomebiology.com in April

RNA transcripts can be sequenced from biological samples, but making sense of those that fail to map exactly to a reference genome is tough. Eric Rivals at the University of Montpellier, France, and his team have written a program called CRAC that can identify tricky transcripts as experimental errors, chromosome rearrangements, small mutations or modifications to messenger RNA. The software simultaneously matches discrete portions of sequenced RNA to locations in the genome and counts up how often unique portions are sequenced — a strategy that combines several computational steps. Although CRAC requires more memory than some similar software, it is more sensitive and precise than other tools for classifying RNA transcripts, the authors say.
Genome Biol. 14, **R30 (2013)**

TING XIN JIANG/SCIENCE AAAS

DEVELOPMENT

Cell signals speckle feathers

A bird's patterned plumage is 'painted' by cell interactions.

Cheng-Ming Chuong at the University of Southern California in Los Angeles and his colleagues found that precursors of pigment-producing cells are positioned in a ring around the base of feather follicles. These precursors divide and develop into pigment-producing cells, which are sent into the feather shaft that emerges from the follicle as feathers grow. Variations in the timing of cell development and positioning of the progenitor cells create distinct designs. Stripes are painted when pigmented cells form in synchronized pulses, and spotted feathers (**pictured**) result from signals that switch pigment synthesis on and off in adjacent, differentiating cells. These subtle modulations allow complex feather patterns to arise during birds' lives and to evolve over time.
Science **http://dx.doi.org/10.1126/science.1230374 (2013)**

CORRECTION

In the Research Highlight 'Mechanics behind sea shell spines' (*Nature* **496,** 9; 2013), the pictured shell belonged to the Strombidae, not Muricidae, family.

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

NIKO BORNEMANN, ALFRED-WEGENER-INSTITUT

Arctic research lab opens for business

A German–Russian expedition team settled into a shiny new research base (pictured) in the Siberian Arctic on 23 April. The Samoylov station, located on a small island in the Lena Delta close to the Laptev Sea, replaces a 15-year-old small wooden station situated nearby. The 14-strong team will study processes affecting the formation and decay of permafrost in the Lena Delta during its four-week stay. Equipped with state-of-the-art lab appliances, the station will enable researchers to conduct fieldwork all year round.

EVENTS

Bomb test detected

The Comprehensive Nuclear-Test-Ban-Treaty Organization (CTBTO) in Vienna said on 23 April that it had detected radioactive gases in the atmosphere at levels indicating that North Korea did test an atomic bomb, as it claimed it had on 12 February. Nuclear monitoring stations in Takasaki, Japan, and in Ussuriysk, Russia, picked up traces of radioactive xenon isotopes — 131m and 133 — which signal a nuclear fission event around the time of the alleged explosion. But the CTBTO says that it is still eliminating other possible explanations, such as releases from a nuclear reactor or other nuclear activity.

POLICY

Helium sales

US legislators voted on 26 April to continue selling federal helium gas reserves. The move follows warnings of a looming shortage in the supply of the gas that researchers and electronics manufacturers use for cooling. The United States was due to stop trading helium reserves in October, once it had paid off debts of $1.3 billion with revenues from the gases' sale. However, the House of Representatives agreed to extend helium sales until all but 85 million cubic metres of the stockpile remain. The Senate is expected to consider a similar proposal on 7 May.

Open discourse

Libel laws that make it harder to suppress free speech in England and Wales came into effect on 25 April, after the Queen gave the Defamation Bill her seal of approval. Scientists have long argued that the previous English libel law threatened open scientific discourse because it favoured those claiming they had been defamed, and that the costs of such cases could force people to keep quiet in the face of legal threats. See go.nature.com/12bxfm for more.

HCFC deal

China will receive up to US$385 million over the next 17 years from the Multilateral Fund for the Implementation of the Montreal Protocol to stop industrial production of hydrochlorofluorocarbons (HCFCs). The ozone-depleting chemicals, which are also powerful greenhouse gases, are used in applications such as refrigeration. The deal, announced on 22 April, makes China party to an existing 2007 global agreement to accelerate the phase out of HCFCs. See go.nature.com/5rzlbk for more.

Pesticide ban

In an effort to protect bees, the European Commission has announced that a two-year ban on the use of three common pesticides on crops will begin on 1 December. The Commission took the decision on 29 April, after a vote by European member states failed to either support or reject the proposed restrictions on the use of neonicotinoids. Scientists argue over whether neonicotinoids damage bee populations. See *Nature* **496,** 408 (2013) and go.nature.com/apvdlf for more.

FACILITIES

Primate pull-out

Harvard Medical School announced on 23 April that it will close its 47-year-old New England Primate Research Center in Southborough, Massachusetts. The centre, which houses 1,860 non-human primates — mostly macaques — will close by 2015 owing to a cash shortage. The animals will be transferred to other primate research centres or be maintained on site, say medical-school officials. Inspections by the US Department of Agriculture found that the centre had violated the Animal Welfare Act several times, with four primate deaths occurring between June 2010 and February 2012. See go.nature.com/zsavjr for more.

Satellite launch

China has launched the first in its series of next-generation civilian Earth-observing

satellites. The Gaofen-1 satellite lifted off on 26 April from the Jiuquan launch facility in the Chinese Gobi desert. Its data will be used to aid distribution of disaster relief and for environmental monitoring, the state-run Xinhua news agency reported. China is planning to launch a further six satellites in the series.

Farewell Herschel

Europe's Herschel mission has come to an end. The €1.1-billion (US$1.4-billion) infrared space telescope exhausted its stores of liquid-helium coolant on 29 April, at which point its scientific instruments stopped working, said the European Space Agency. Astronomers have hailed the legacy of the observatory, which over three years has helped them to revise theories about the birth and death of stars (see *Nature* **495**, 151–152; 2013).

RESEARCH

Freshwater lifeline

Ontario's government threw a lifeline to Canada's Experimental Lakes Area (ELA) on 24 April. Funding shortages led the Canadian government to close the freshwater research facility (**pictured**) in March. Ontario's premier, Kathleen Wynne, says that the province will provide funds to support the facility

HANNAH HOAG

and will work towards a deal to transfer research operations to the International Institute for Sustainable Development, a think tank based in Winnipeg, Manitoba. See go.nature.com/q39xpw for more.

Hepatitis drug

A new hepatitis-C drug, sofosbuvir, has been found to be highly effective in clinical trials. Developed by Gilead of Foster City, California, the drug is one of several in development that could replace existing hepatitis-C treatments, which can cause harsh side effects. See go.nature.com/fpasug and page 18 for more.

GM salmon

Genetically modified (GM) salmon have moved one step closer to US grocery stores. On 26 April, the US Food and Drug Administration (FDA) closed a public consultation on its finding that the engineered fish pose no significant environmental concern. The FDA must now evaluate the comments before finalizing its decision. See page 17 for more.

PEOPLE

Research politics

Physicist Maria Chiara Carrozza was appointed as research minister in Italy's new government on 27 April. Carrozza is a biorobotics specialist and was rector of an elite university in Pisa, the Sant'Anna School of Advanced Studies, until she was elected to parliament in February for the centre-left Democratic Party. Carrozza says that she will boost Italy's notoriously low spending on science, increase the number of research and academic positions and reduce the bureaucratic red tape that encumbers research.

Lab-death trial

Patrick Harran, a chemist at the University of California, Los Angeles, will stand trial for the death of research assistant Sheharbano Sangji, who died more than four years ago in a laboratory accident. Harran will be tried on three counts of violating health and safety standards, a Los Angeles judge ordered on 26 April. A trial date has not yet been set; Harran could face 4.5 years in prison. The case is touted as the first time that a scientist has gone to trial over an accident in a US academic lab. See go.nature.com/738hpf for more.

Wellcome head

Jeremy Farrar, an expert in infectious diseases, is the next director of the Wellcome Trust, the London-based biomedical research charity announced on 24 April. Farrar will take up the post in October succeeding Mark Walport, who left in March after a decade in the job to become the UK government's chief scientific adviser. See page 19 for more.

Jailed physicist

Omid Kokabee, a former physics graduate student who has been imprisoned in Iran since early 2011, has written a public letter stating that he was jailed for refusing to cooperate with Iranian military projects. The letter, dated March 2013, was revealed by *Nature* last week. In a separate private letter, Kokabee, who had been studying laser physics at the University of Texas at Austin, claims that his expertise was sought for nuclear applications. See go.nature.com/4mught for more.

COMING UP

5–8 MAY
The third World Conference on Research Integrity meets in Montreal, Canada.
go.nature.com/enai63

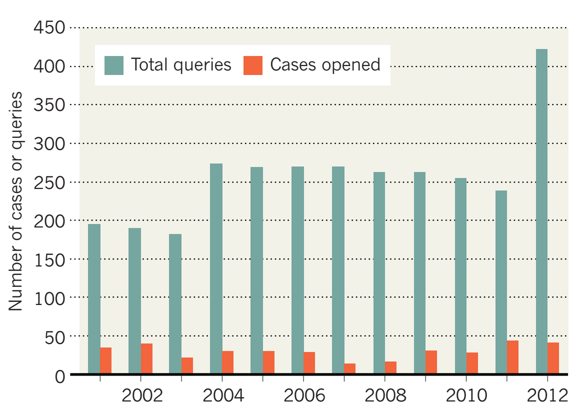
6–8 MAY
Darmstadt, Germany, hosts the first international meeting on research into 'ocean colour' science: how satellite observations of the ocean can infer photosynthesis and other activity from colour.
go.nature.com/ubntid

TREND WATCH

The office that oversees misconduct investigations involving US-government-funded biomedical researchers has seen the number of allegations it has received since 2001 rise by 216%. Officials at the Office of Research Integrity in Rockville, Maryland, said in April that the office is likely to receive more than 500 queries in 2013. But because the number of staff who process allegations has remained flat, at 8–10 people, the office cannot take on any more cases.

MISCONDUCT OFFICE OVERLOADED

The US Office of Research Integrity saw a jump in queries last year — although limited staff numbers restrict its capacity to open new cases.



SOURCE: ORI

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

RENE CLEMENT/POLARIS/EYEVINE

Continuing reliance on coal, which fuels this power plant in Germany, is driving carbon dioxide levels in the atmosphere ever higher.

CLIMATE

Global carbon dioxide levels near worrisome milestone

Concentrations of greenhouse gas will soon surpass 400 parts per million at sentinel spot.

BY RICHARD MONASTERSKY

Near the moonscape summit of the Mauna Loa volcano in Hawaii, an infrared analyser will soon make history. Sometime in the next month, it is expected to record a daily concentration of carbon dioxide in the atmosphere of more than 400 parts per million (p.p.m.), a value not reached at this key surveillance point for a few million years.

There will be no balloons or noisemakers to celebrate the event. Researchers who monitor greenhouse gases will regard it more as a disturbing marker of humanity's power to alter the chemistry of the atmosphere and by extension, the climate of the planet. At 400 p.p.m., nations will have a difficult time keeping global warming in check, says Corinne Le Quéré, a climate researcher at the University of East Anglia in Norwich, UK, who says that the impact "is getting very dangerously close to reaching the 2 °C target that governments around the world have pledged not to exceed".

It will be a while, perhaps a few years, before the global CO_2 concentration averaged over an entire year, passes 400 p.p.m.. But topping that value at Mauna Loa is significant because researchers have been monitoring the gas there since 1958, longer than any other spot. "It's a time to take stock of where we are and where we're going," says Ralph Keeling, a geochemist at the Scripps Institution of Oceanography in La Jolla, California, who oversees that centre's CO_2 monitoring efforts on Mauna Loa. That gas record, known as the Keeling curve, was started by his father, Charles Keeling.

When monitoring started, the CO_2 level stood at 316 p.p.m., not much higher than the 280 p.p.m. that characterized conditions ▶

before the industrial revolution. But since the Hawaiian measurements began, the values have followed an upward slope that shows no sign of levelling off (see 'On the rise'). Emissions of other greenhouse gases are also increasing, pushing the total equivalent concentration of CO_2 in the atmosphere to around 478 p.p.m. in April, according to Ronald Prinn, an atmospheric scientist at the Massachusetts Institute of Technology in Cambridge.

Data compiled by Le Quéré and other members of the Global Carbon Project suggest that humans contributed around 10.4 billion tonnes of carbon into the atmosphere in 2011. About half of that is taken up each year by carbon 'sinks' such as the ocean and vegetation on land; the rest remains in the atmosphere and raises the global concentration of CO_2.

"The real question now is: how will the sinks behave in the future?" says Gregg Marland, an environmental scientist at Appalachian State University in Boone, North Carolina, who helps to compile the emissions data.

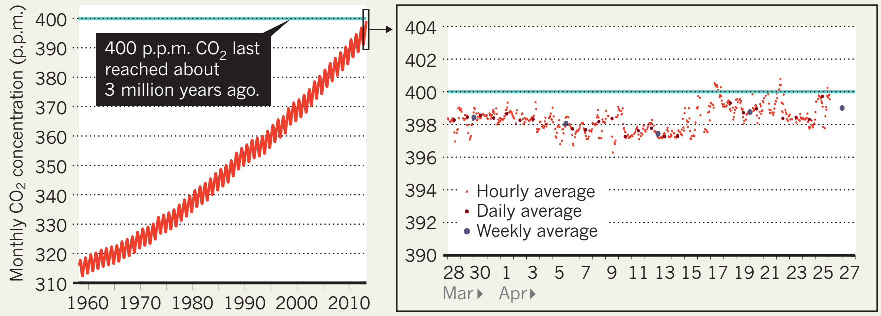
The sinks have grown substantially since Keeling began his measurements, when carbon emissions totalled about 2.5 billion tonnes a year. But climate models suggest that the land and ocean will not keep pace for long.

"At some point the planet can't keep doing us a favour, particularly the terrestrial biosphere," says Jim White, a biogeochemist at the University of Colorado Boulder. As the sinks slow down and more emitted CO_2 stays in the atmosphere, levels will rise even faster.

Some researchers have suggested that the sinks have already started to clog up, reducing their ability to take up more CO_2 (J. G. Canadell *et al. Proc. Natl Acad. Sci. USA* **104,** 18866–18870; 2007). Others disagree.

ON THE RISE

Measurements of atmospheric CO_2 levels at Mauna Loa, Hawaii, show that the greenhouse gas has accumulated steadily, and spiked above 400 parts per million (p.p.m) several times in April.



SOURCE: SCRIPPS INSTITUTION OF OCEANOGRAPHY AT UC SAN DIEGO

Ashley Ballantyne, a biogeochemist at the University of Montana in Missoula, worked with White and others to examine records of emissions as well as CO_2 measurements made around the globe. They found no signs of sinks slowing down (A. P. Ballantyne *et al. Nature* **488,** 70–72; 2012). But it is difficult to be sure, says Inez Fung, a climate modeller at the University of California, Berkeley. "We don't have adequate observing networks." The largest global network, operated by the US National Oceanic and Atmospheric Administration, had to trim 12 stations in 2012 because of budget cuts.

Some of the most crucial areas, such as the tropics, are also the least monitored, although researchers are seeking to fill in the gaps. Scientists from Germany and Brazil are building a 300-metre tower to keep tabs on the Amazon (see *Nature* **467,** 386–387; 2010). And Europe's Integrated Carbon Observation System is setting up stations throughout the continent and at some marine sites to measure CO_2 and other greenhouse gases.

Satellites, too, could monitor carbon sources and sinks. Two orbiters are already providing some data, and NASA plans to launch the much anticipated Orbiting Carbon Observatory-2 next year (see page 5). An earlier version of that satellite failed during its 2009 launch.

Even as new resources come online, however, researchers are struggling to keep the Mauna Loa station going. "The amount of money that I'm able to obtain for the programme has diminished over time," says Keeling, whose group monitors CO_2 concentration at 13 sites around the world.

"It's kind of silly that we chose to go all ostrich-like," says White of the funding difficulties. "We don't want to know how much CO_2 is in the atmosphere, when we ought to be monitoring even more." ■

GENETICS

Flu papers spark row over credit for data

Rush to publish on H7N9 avian flu upsets Chinese scientists.

BY DECLAN BUTLER AND DAVID CYRANOSKI

On 31 March, China reported the first human cases of infection with a new H7N9 avian flu virus. The same day, a team at the Chinese National Influenza Center (CNIC) in Beijing uploaded to a research database the genetic sequences of viruses isolated from the first three human cases. But *Nature* has learned that in the days that followed, Chinese scientists and officials grew increasingly concerned that China might lose credit for its work in isolating and sequencing the virus.

The sequences were placed in the Global Initiative on Sharing All Influenza Data (GISAID) database. According to the database's rules, scientists who use sequences from it must credit those who deposited the data and, where possible, propose collaborations with them.

"Unfortunately some bad things happened when we released the sequences in GISAID, and they really hurt us," says Yuelong Shu, head of the CNIC, which is also the World Health Organization (WHO) Collaborating Centre for Reference and Research on Influenza in China. "GISAID have tried their best to help us," he adds. "I really appreciate what they have done."

Shu did not initially reveal specific concerns, but other researchers have told *Nature* some of the details. On 5 April, the Chinese scientists submitted their first major H7N9 paper, including analyses of the sequences, to *The New England Journal of Medicine* (*NEJM*).

Around the same time, the researchers learned that they might be scooped: several other research groups were preparing to publish papers on the virus, or already had done so, including analyses of the sequences in GISAID.

This news was followed by what seemed to be a snub. It emerged on 5 April that drug firm Novartis in Basel, Switzerland, and the J. Craig Venter Institute in Rockville, Maryland, planned to use the uploaded sequences to develop H7N9 vaccines. The initiative had US government funding and would be a collaboration with the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia — but not with the Chinese team. The Chinese researchers felt that this was not in the spirit of GISAID.

The sharing of flu-virus data and materials

CHINAFOTOPRESS/PHOTOSHOT

Chinese medical workers take part in a drill to simulate an outbreak of H7N9 flu in the population.

FLU TRACKING

A virus on the move

As the H7N9 avian influenza outbreak in China enters its fourth week, the virus has expanded it geographic range. Human cases have been reported in Fujian province, hundreds of kilometres south of the main outbreak area around Shanghai, and in the Jiangxi and Hunan provinces at similar distances to the southwest.

As of 29 April, the World Health Organization (WHO) had confirmed 126 cases, including 24 deaths, up from 104 confirmed cases on 22 April.

On 24 April, a China–WHO Joint Mission including a team of international flu experts ended a week-long investigation of the outbreak. They noted that some family clusters of cases have occurred, which could signal either limited human-to-human spread or infection from a common source. But for the moment, there is no evidence of sustained human-to-human spread.

On the same day, the first human case outside mainland China was reported in Taiwan, but the 53-year-old man is thought to have caught the disease while on the mainland.

The outbreak is "complex and difficult and is evolving", says Keiji Fukuda, the WHO's assistant director-general for health security. The most probable source of the human infections is birds, particularly poultry, at live poultry markets, he says. **D.B.**

has long been a politically charged issue in global health. Timely information from potentially pandemic flu strains is crucial for efforts to monitor drug resistance and the evolution of viruses, and for the development of diagnostics and vaccines. But some countries have been reluctant to share such data because they have seen little in return in terms of collaboration, technology transfer or access to the drugs and vaccines developed as a result.

GISAID was created in 2008 to help overcome some of these concerns. "Without a mechanism like GISAID it would be very difficult for various authorities to make information available prior to publication," says Alan Hay, co-chair of GISAID's scientific advisory council.

Novartis spokeswoman Liz Power says that after the company's researchers downloaded the H7N9 sequences, it "explored research collaboration" with the Chinese CDC in Beijing, of which the CNIC is part. "We are committed to sharing any meaningful insights coming out of our work with China CDC," she says.

Kristine Sheedy, a spokeswoman for the US CDC, acknowledges that "there were differences in understanding and expectations regarding use of the Chinese H7N9 sequence data by several outside groups", but adds that the US CDC was not among them. The CDC has had a strong ongoing collaboration with its counterpart in China since the start of the H7N9 outbreaks, she says.

Shu says that the Chinese researchers would have preferred for the vaccine developers to have told them in advance about how they intended to use the sequences, but adds that communication channels have now been opened and that the various parties have agreed to collaborate. "Thanks to the president of GISAID this situation was quickly mitigated," says Shu.

Chinese worries over being scooped also seem to have been put to rest. The CNIC scientists were most concerned about a major analysis of the H7N9 virus scheduled for publication on 10 April in *Eurosurveillance* — which would have appeared before the Chinese *NEJM* paper.

A co-author on the *Eurosurveillance* paper, virologist Masato Tashiro of the Influenza Virus Research Center in Tokyo, the WHO's influenza reference centre in Japan, says that he sent a draft of the paper to the Chinese researchers on 7 April, inviting them to be co-authors. They declined, but asked Tashiro to delay publication until after their *NEJM* paper had appeared. He agreed and the *NEJM* paper was published on 11 April (R. Gao *et al. N. Engl. J. Med.* http://doi.org/k7r; 2013), with the *Eurosurveillance* paper appearing later the same day (T. Kageyama *et al. Euro Surveill.* **18,** 20453; 2013). Tashiro notes that holding the paper did not have an impact on public health, because all its analyses were shared with the WHO's global network of flu labs on 1 and 2 April, and were used to help the WHO to prepare its initial risk assessment of the virus (see 'Flu tracking').

"One has to recognize the sensitivities in relation to scientific priority," says Hay, who thinks that many potential difficulties could be avoided if people spoke to each other more about their work and their publication plans.

"Scientific etiquette is without doubt a key to keeping the rapid sharing of data a reality," says Shu. In this case, he continues, "after some initial concerns we found that both researchers and publishers were understanding of our predicament".

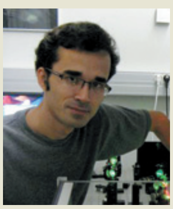
Hay hopes that the hiccups won't discourage Chinese researchers from making their H7N9 data publicly accessible quickly. "It is very important to continue to share sequences from the more recent cases," he says.

For his part, Shu says that he is keen to ensure that researchers continue to have "unfettered access to data". ■

MORE ONLINE

O. KOKABEE

TOP STORY

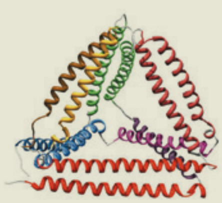


Iranian physicist says he was jailed for refusing to do military research. go.nature.com/4mught

MORE NEWS

- Ant family tree validates competing biodiversity theories go.nature.com/6r3uiw
- Social networks can help in manhunts if used carefully go.nature.com/mpybwy
- Google Trends predicts market — but only in hindsight go.nature.com/nxywlw

IMAGE OF THE WEEK



Engineered protein gets in on DNA's origami act go.nature.com/v6qf4c

PSYCHOLOGY

Disputed results a fresh blow for social psychology

Failure to replicate intelligence-priming effects ignites row in research community.

BY ALISON ABBOTT

Thinking about a professor just before you take an intelligence test makes you perform better than if you think about football hooligans. Or does it? An influential theory that certain behaviour can be modified by unconscious cues is under serious attack.

A paper published in *PLoS ONE* last week[1] reports that nine different experiments failed to replicate this example of 'intelligence priming', first described in 1998 (ref. 2) by Ap Dijksterhuis, a social psychologist at Radboud University Nijmegen in the Netherlands, and now included in textbooks.

David Shanks, a cognitive psychologist at University College London, UK, and first author of the paper in *PLoS ONE*, is among sceptical scientists calling for Dijksterhuis to design a detailed experimental protocol to be carried out in different laboratories to pin down the effect. Dijksterhuis has rejected the request, saying that he "stands by the general effect" and blames the failure to replicate on "poor experiments".

An acrimonious e-mail debate on the subject has been dividing psychologists, who are already jittery about other recent exposures of irreproducible results (see *Nature* **485,** 298–300; 2012). "It's about more than just replicating results from one paper," says Shanks, who circulated a draft of his study in October; the failed replications call into question the underpinnings of 'unconscious-thought theory'.

> ***"It's about more than just replicating the results from one paper."***

Dijksterhuis published that theory in 2006 (ref. 3). It fleshed out more general, long-held claims about a 'smart unconscious' that had been proposed over the past couple of decades — exemplified in writer Malcolm Gladwell's best-selling book *Blink* (Penguin, 2005). The theory holds that behaviour can be influenced, or 'primed', by thoughts or motives triggered unconsciously — in the case of intelligence priming, by the stereotype of a clever professor or a stupid hooligan. Most psychologists accept that such priming can occur consciously, but many, including Shanks, are unconvinced by claims of unconscious effects.

Social psychologist Ap Dijksterhuis.

FLIP FRANSSEN/HOLLANDSE HOOGTE/EYEVINE

In their paper, Shanks and his colleagues tried to obtain an intelligence-priming effect, following protocols in Dijksterhuis's papers or refining them to amplify any theoretical effect (for example, by using a test of analytical thinking instead of general knowledge). They also repeated intelligence-priming studies from independent labs. They failed to find any of the described priming effects in their experiments.

The e-mail debate that Shanks joined was kicked off last September, when Daniel Kahneman, a Nobel-prizewinning psychologist from Princeton University in New Jersey who thinks that unconscious social priming is likely to be real, circulated an open letter warning of a "train wreck looming" (see *Nature* http://doi.org/mdr; 2012) because of a growing number of failures to replicate results. Social psychology "is now the poster child for doubts about the integrity of psychological research", he told psychologists, "and it is your responsibility" to deal with it.

Other high-profile social psychologists whose papers have been disputed in the past two years include John Bargh from Yale University in New Haven, Connecticut. His claims include that people walk more slowly if they are primed with age-related words.

Bargh, Dijksterhuis and their supporters argue that social-priming results are hard to replicate because the slightest change in conditions can affect the outcome. "There are moderators that we are unaware of," says Dijksterhuis.

But Hal Pashler, a cognitive psychologist at the University of California, San Diego — a long-time critic of social priming — notes that the effects reported in the original papers were huge. "If effects were that strong, it is unlikely they would abruptly disappear with subtle changes in procedure," he says.

No one is suggesting that there is anything fraudulent about the results, but the charges that some of Dijksterhuis's key papers may report false positives is a particular embarrassment for the Netherlands. It comes close on the heels of exposures of scientific misconduct by two other Dutch social psychologists: in 2011, Diederik Stapel of Tilburg University admitted to inventing data, and in June 2012, an investigation committee concluded that Dirk Smeesters from the Erasmus University in Rotterdam had cherry-picked data in some papers.

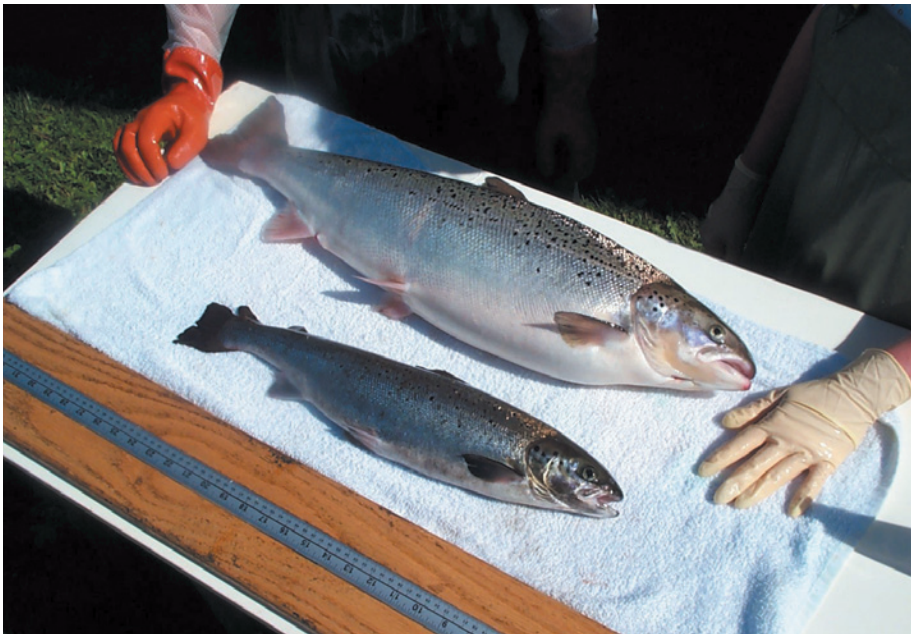
Shanks's replication failures cannot be dismissed, says Eric-Jan Wagenmakers, a mathematical psychologist at the University of Amsterdam who last year published a series of studies that failed to lend support[4] to unconscious-thought theory. He is disappointed that Dijksterhuis has declined "repeated requests" to help to generate a definitive answer.

Dijksterhuis says that "focusing on a single phenomenon is not that helpful and won't solve the problem". He adds that social psychology needs to get more rigorous, but that the rigour should be applied to future, not historical, experiments. The social-priming debate will rumble on, he says, because "there is an ideology out there that doesn't want to believe that our behaviour can be cued by the environment".

Others remain concerned. Kahneman wrote in the e-mail debate on 4 February that this "refusal to engage in a legitimate scientific conversation … invites the interpretation that the believers are afraid of the outcome". ■

1. Shanks, D. R. *et al. PLoS ONE* **8,** e56515 (2013).
2. Dijksterhuis, A. & van Knippenberg, A. *J. Pers. Soc. Psychol.* **74,** 865–877 (1998).
3. Dijksterhuis, A. & Nordgren, L. F. *Persp. Psychol. Sci.* **1,** 95–109 (2006).
4. Huizenga, H. M., Wetzels, R., van Ravenzwaaji, D. & Wagenmakers, E.-J. *Organ. Behav. Hum. Decis. Proc.* **117,** 332–340 (2012).

AQUABOUNTY



A genetically engineered salmon (top) grows twice as fast as its wild counterpart (bottom).

BIOTECHNOLOGY

Transgenic salmon nears approval

Slow US regulatory process highlights hurdles of getting engineered food animals to dinner tables.

BY HEIDI LEDFORD

In the remote highlands of Panama, in tanks protected by netting, barbed wire and guard dogs, swim the world's most expensive and scrutinized fish. These swift-growing salmon have been at the centre of a 18-year, US$60-million battle to bring the first genetically modified (GM) animal to US dinner tables — a struggle that may be nearing its end.

Last week marked the end of the public's opportunity to weigh in on a US Food and Drug Administration (FDA) draft assessment of the salmon. Genetically engineered to grow twice as fast as their unaltered brethren, the fish pose no significant environmental threat to the United States when grown in landlocked tanks, says the FDA. The agency needs only to finalize that assessment before deciding whether to approve the fish for human consumption. The number of opportunities for a surprise delay — a recurring theme in the history of these salmon — is dwindling (see 'Against the current').

Environmental groups are preparing to take the battle to consumers by fighting the sale of the fish in grocery stores across the country. Others point out that it will be years before the salmon are anything more than a curiosity. At full capacity, the Panama facility can produce only about 100 tonnes of salmon a year, says Gregory Jaffe, director of biotechnology at the Center for Science in the Public Interest, a consumer group in Washington DC that monitors the regulation of GM foods. That amount is a trifle compared to the roughly 230,000 tonnes of farmed Atlantic salmon that the United States imported in 2012. "You'd have to try hard to eat it," says Jaffe. "It won't be as hard as winning the lottery, but it will be close."

For the firm that developed the fish, AquaBounty Technologies of Maynard, Massachusetts, those 100 tonnes are a hard-won prize. In 1989, the salmon were engineered to overexpress a growth-hormone gene. The result: 'AquAdvantage' fish that grew to full size in around 18 months rather than the usual 3 years. The company applied for FDA approval in 1995 and has been stuck in regulatory limbo ever since. AquaBounty has had to demonstrate the food's safety, and gauge the environmental risk of the sterile fish escaping its tanks and successfully mating with wild salmon. By contrast, the FDA approved the first GM crop for human consumption — the Flavr Savr tomato — after just three years of regulatory consideration.

CASH CRISIS

The uncertainty has taken its toll. To save money, AquaBounty has reduced its staff by more than half. Last year, the company sold off its research and development arm and lost one of its biggest investors. In March, AquaBounty came within a week of running out of cash, says chief executive Ronald Stotish. The firm was saved by last-minute refinancing and fresh investment from Intrexon, a synthetic-biology company based in Blacksburg, Virginia.

At first glance, the Panama facility hardly seems to be the key to financial prosperity. With salmon selling for around $6.50 per kilogram, AquaBounty would make less than $1 million each year from the salmon. It would take decades for the company to make back its $60-million investment if it relied solely on the Panama farm.

Stotish says that the company must expand. Following FDA approval, AquaBounty hopes to sell its salmon eggs to farmers and expand to markets in Argentina, Canada, Chile and China.

To sell AquAdvantage fish in the United States, each farm would require separate FDA approval, but because the food safety of the fish has already been vetted, the approval process would require only an environmental evaluation, says Jaffe.

Yet even with regulatory approval, the battle over AquaBounty's salmon will be far from over. In March, several speciality grocery stores, including Whole Foods, an international chain based in Austin, Texas, said that they would not sell AquAdvantage fish. Lawmakers in Alaska and Oregon, which both export wild salmon, have repeatedly tried to block the GM fish because they fear contamination of the wild stock and worry that it could drive down the price of farmed salmon.

AquaBounty's long struggle has discouraged other US companies from producing GM animals for food. Mark Walton, chief marketing officer at Recombinetics, an animal-biotechnology company in St Paul, Minnesota, says that his company will focus initially on medical applications — using modified farm animals as disease models, for example — rather than on livestock for food. Medical applications of GM technology do not stir consumer passions in the same way as GM foods, and there is a regulatory precedent: in 2009, the FDA approved a goat that makes an anti-clotting drug in its milk. If Recombinetics invests in agricultural products, Walton adds, the items will ▶

AGAINST THE CURRENT

The US Food and Drug Administration (FDA) has been slow to approve a genetically modified (GM) salmon made by AquaBounty of Maynard, Massachusetts. The fish would be the first GM animal authorized for human consumption.

AQUABOUNTY

1989 Canadian researchers engineer wild Atlantic salmon to overexpress growth hormone.

1995 AquaBounty files an Investigational New Animal Drug application with the FDA.

2001 AquaBounty submits its first regulatory study to the FDA.

2009 The FDA releases guidance for its evaluation of genetically engineered animals as veterinary drugs; AquaBounty completes its FDA submission.

2010 The FDA says that GM salmon is safe to eat.

2012 The FDA completes its draft environmental assessment in May, but does not release it to the public until December.

2013 The public-comment period for the draft environmental assessment is extended by two months and concludes on 26 April.

▸ probably be marketed outside the United States first. "The AquaBounty example has [made] the company very sceptical about how much investment to pour into the US regulatory process," he says.

Yet Stotish says that GM animal products will inevitably find their way to grocery stores. He points to heavy investment in the technology in China, where dozens of GM farm animals are in development. "I think we will end up eating genetically modified animals of a variety of species," says Stotish. "But they'll come from other countries." ■

MEDICINE

Targeted drugs to tackle hepatitis C

But experts debate US screening recommendations.

BY BETH MOLE

John strains to recall the gap between learning that he had hepatitis C and deciding to get treated: it was either four years or five. His thinking is clouded by the combination of three drugs that he is taking to clear the infection. After the treatments' other side effects set in — severe flu-like symptoms, depression and exhaustion — he took leave from his job as a chef in New York. John, whose name has been changed to protect his privacy, was at high risk of catching the virus, having once been addicted to crystal methamphetamine. But as a 51-year-old, he is also a baby boomer — a member of the generation born between 1945 and 1965 — millions of whom will face the disease and its sometimes harrowing treatment.

Better drugs are on the way. But the possibility of improved treatment is intensifying a debate about whether to screen a broad swathe of the US population for hepatitis C.

Last month, the pharmaceutical company Gilead, based in Foster City, California, submitted its hepatitis-C drug sofosbuvir to the US Food and Drug Administration for approval, after phase II trials showed a 100% success rate in a few patient groups when it was used in combination with existing drugs. Last week, the first phase III results showed similarly promising results (E. Lawitz *et al. N. Engl. J. Med.* http://doi.org/mcc; 2013).

The drug is one of at least ten in phase III trials in the United States that promise to improve results or reduce side effects. The first of these drugs could reach the market as early as 2014, and a recommendation from the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, to screen an entire generation for the disease could create vast demand for them.

John is a part of a demographic time bomb. Up to 4 million Americans are infected with hepatitis C, which can irreparably damage the liver and lead to liver cancer, but because it inflicts injury slowly over decades, as many as 85% of carriers do not know that they have it. Baby boomers account for about 27% of the US population, but up to 75% of those infected with hepatitis C, possibly because injecting drugs — one infection route — was more common during their youth than in other eras. Last August, the CDC recommended screening the entire generation of people born between 1945 and 1965, as well as people in high-risk populations such as intravenous-drug users. The CDC predicts that generational screening would find an extra 800,000 cases and prevent at least 120,000 deaths. "We have an opportunity to make a real dent in the impact of the disease," says Kimberly Page, an epidemiologist at the University of California, San Francisco.

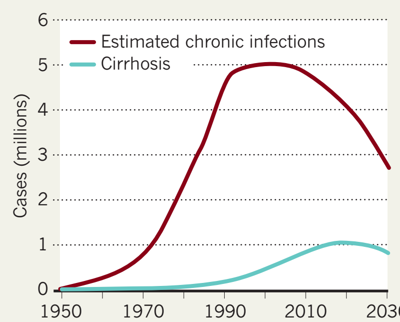
John's doctor, infectious-disease specialist Kristen Marks of Weill Cornell Medical College in New York, says that screening is especially important for baby boomers because early symptoms of hepatitis C, such as fatigue and malaise, are difficult to distinguish from signs of ageing. People dismiss symptoms, says Marks, and some might not remember trying intravenous drugs in their youth. Even if they do, she adds, "they might not tell their doctor". A peak in cases of liver scarring from untreated hepatitis C is expected in the next few years (see 'An approaching burden'). But with the new drugs on the horizon, now is an optimistic time for treatment, says Marks. "Historically, not having good treatments was a disincentive for screening," she says. "Now, I think there's a renewed interest."

But last November, the US Preventive Services Task Force (USPSTF), a panel of experts assembled by the US Department of Health and Human Services, released a draft statement giving the screening recommendation a 'grade C'.

➲ NATURE.COM
For more on Hepatitis C, visit:
go.nature.com/yxxtwh

AN APPROACHING BURDEN

The high number of hepatitis-C infections in the United States is expected to lead to a peak in cases of cirrhosis, or liver scarring, by around 2020.



SOURCE: G. L. DAVIS *ET AL. GASTROENTEROLOGY* **138**, 513–521 (2010)

That means that doctors should consider birth year when deciding whether to offer screening, but should take other factors into account. The mediocre grade could discourage many health-care providers — including Medicaid, the provider for people with low incomes — from pushing screenings.

As with its controversial recommendations in 2009 and 2012 to limit screening for breast and prostrate cancer, the USPSTF has tried to balance the benefits of screening against its costs and the risk of unnecessary treatment. The combination therapies used to combat hepatitis C can cost US$1,100 per week and last for up to a year, with severe side effects. Other treatments cost $4,100 per week. (Gilead declined to comment on the future price of sofosbuvir-based treatments.)

Roger Chou, an internal-medicine specialist at Oregon Health and Science University in Portland and a scientific reviewer for the USPSTF, adds that in most patients, the disease is imperceptible: only 20% of people develop liver scarring in the first 20 years of infection, according to the CDC. Of the few baby boomers that might be caught through additional screening, says Chou, some will not need to be treated.

> ***"We have an opportunity to make a real dent in the impact of the disease."***

But new drugs, however expensive, could change the calculus for doctors and patients, says Mark Eckman, a physician at the University of Cincinnati in Ohio, who has calculated that even screening the entire US population would be cost effective given the financial and personal burdens of living with liver diseases (M. H. Eckman *et al. Clin. Infect. Dis.* **56,** 1382–1393; 2013).

For example, sofosbuvir, which is one of a set of new antiviral drugs that specifically target hepatitis C rather than viruses in general, can achieve success rates above 90% in combination treatments of just three months. The drug inhibits the virus's RNA polymerase, preventing viral replication. It is also being tested without the classic combination drug of pegylated interferon, which boosts the immune system but causes harsh side effects.

The USPSTF is still reviewing its draft recommendations, but it is likely to make a final decision in the next few months, well before approval of sofosbuvir or other new drugs could alter the calculations.

That is too bad, says David Thomas, a viral-hepatitis specialist at Johns Hopkins University in Baltimore, Maryland, who argues that the next generation of drugs helps to justify wide-scale screening. "It makes a pretty easy case for doing something different," he says. ■

BIOMEDICINE

Clinician to head Wellcome Trust

Jeremy Farrar to lead one of world's largest research charities.

BY RICHARD VAN NOORDEN

From his base in Vietnam, Jeremy Farrar has spent the past 17 years on the front line of the battle with infectious diseases, from dengue and typhoid to severe acute respiratory syndrome (SARS) and now H7N9 avian influenza. The British clinician has led the Oxford University Clinical Research Unit in Ho Chi Minh City as it has grown from a dozen people to around 200 researchers supporting public-health efforts in Vietnam, Nepal and Indonesia.

Now, Farrar is stepping up to lead the UK institution that paid for much of his work: the Wellcome Trust, one of the world's largest charities funding biomedical research. Colleagues and public-health leaders say that the trust, which last year spent £746 million (US$1.15 billion), has made an excellent choice — and wonder whether it signals an even greater focus on funding research in developing countries.

"He's massively driven, and a great visionary. He's invested his career in doing the research where the problem lies; he believes tropical medicine should be done in the tropics," says Bob Snow, one of Farrar's collaborators, who works on malaria and public health in Nairobi.

Farrar moved to Vietnam in 1996, when the Wellcome Trust was boosting investment in disease-ridden countries in Africa and southeast Asia. He saw the SARS outbreak in 2003 at close quarters — his friend, Carlo Urbani, died of the virus while working for the World Health Organization (WHO) in Hanoi. Then came a surge in H5N1 avian flu, which hit Vietnam hard. "It was a tense time for everyone," says Cameron Simmons, a dengue expert who works with Farrar; clinicians were treating patients and trying to explain the crisis. Through all of this, Farrar's leadership and ability to build trust between people was evident, says Simmons.

"Jeremy's very much a shrewd team player," says Colin Blakemore, a neuroscientist at the University of Oxford, UK, and former head of Britain's Medical Research Council. Farrar has brokered funding from several sources, and his centre's work on flu required negotiations with countries such as China to obtain samples. Those skills will serve him well when he moves to the Wellcome Trust in October.

For the past decade, Farrar has been migrating to a more strategic role, Snow says, serving on WHO advisory boards and pushing for a greater focus on flu surveillance and on capacity-building in the developing world.

WELLCOME LIBRARY, LONDON

Clinician Jeremy Farrar.

"I believe that we have to bring some of the huge investment by the developed world in genomics, technology and training to affected countries in Asia and elsewhere," Farrar wrote last year (*Nature* **483,** 534–535; 2012).

Farrar would not divulge whether his vision of international public-health strategy would affect the trust's priorities: "I have too much to do on H7N9 and hand, foot and mouth disease to talk about that," he says. But Snow says that researchers are hopeful. Since 2008, the charity — under its previous director Mark Walport, now the UK government's chief scientific adviser — has increased its spending outside the United Kingdom from 14% to 22%, expanding support for programmes in India and sub-Saharan Africa in particular. It has also doubled the share of its cash it gives to translational research, from 6% to 12%. David Heymann, chairman of the advisory board for Public Health England, says that Farrar is likely to encourage those trends.

"It wasn't an easy decision for him to give up the science," says epidemiologist Simon Hay, a collaborator at the University of Oxford, "but there's a responsibility for people that have been advocating in public health to step up when these positions come up." Moreover, the Vietnam unit can now operate without Farrar.

"There will be tears from our end and from our Vietnamese partners. The trust is very lucky to be getting Jeremy — he's a remarkable leader," says Simmons. ■

Additional reporting by Ewen Callaway and David Cyranoski.

CORRECTION

The News Feature 'The gun fighter' (*Nature* **496,** 412–415; 2013) did not cite the sources for the graphs. These have now been added online.

SPECIAL ISSUE

TARNISHED PROMISE

Genetically modified crops generate hype and hatred. A special section of Nature *cuts through the drama.*

Foreign genes were successfully introduced into plants for the first time 30 years ago (see page 40). Ever since, genetically modified (GM) crops have promised to deliver a second green revolution: a wealth of enhanced foods, fuels and fibres that would feed the starving, deliver profits to farmers and promote a greener environment. In many ways, that revolution has arrived. Crops engineered to carry useful traits now grow on 170 million hectares in at least 28 countries (see page 22).

But to many, GM crops have been a failure. The market is dominated by just a few insect-resistant and herbicide-tolerant crops. The environmental benefits are disputed, and activists question the safety of GM foods. Politicized and polarized, the war of words that surrounds GM crops ignores the complex truths.

In this special issue, *Nature* explores the messy middle ground. A News Feature weighs the evidence behind some of the most controversial claims about the effects of GM crops (see page 24). Christopher Whitty, chief scientific adviser at the UK Department for International Development, and his colleagues argue that the negative attitudes towards GM crops in the developed world undermine the technology's potential in the developing one (see page 31). Such sentiments have helped to delay the approval of the first genetically modified animal for human consumption, a fast-growing salmon (see page 17).

The next generation of GM crops might benefit from these hard lessons. Fusuo Zhang, director of the Center for Resources, Environment and Food Security at China Agricultural University in Beijing, thinks that his country — now the sixth-largest adopter of GM crops — will serve as a hothouse for agricultural technologies (see page 33). A Perspective article reviews research on membrane transporters in plants that could lead to traits such as stress resistance and increased nutrient content (see page 60). And a second News Feature explores the genetic engineering technologies that are giving rise to the next generation of GM crops (see page 27). The battles are by no means over, but the hope is that science and reasoned debate can inform the future of these technologies. ■

IMAGE: KELLY KRAUSE/*NATURE* (PHOTO: NAGY-BAGOLY ARPAD/SHUTTERSTOCK)

GM CROPS: PROMISE & REALITY
A *Nature* special issue
nature.com/gmcrops

SHUTTERSTOCK

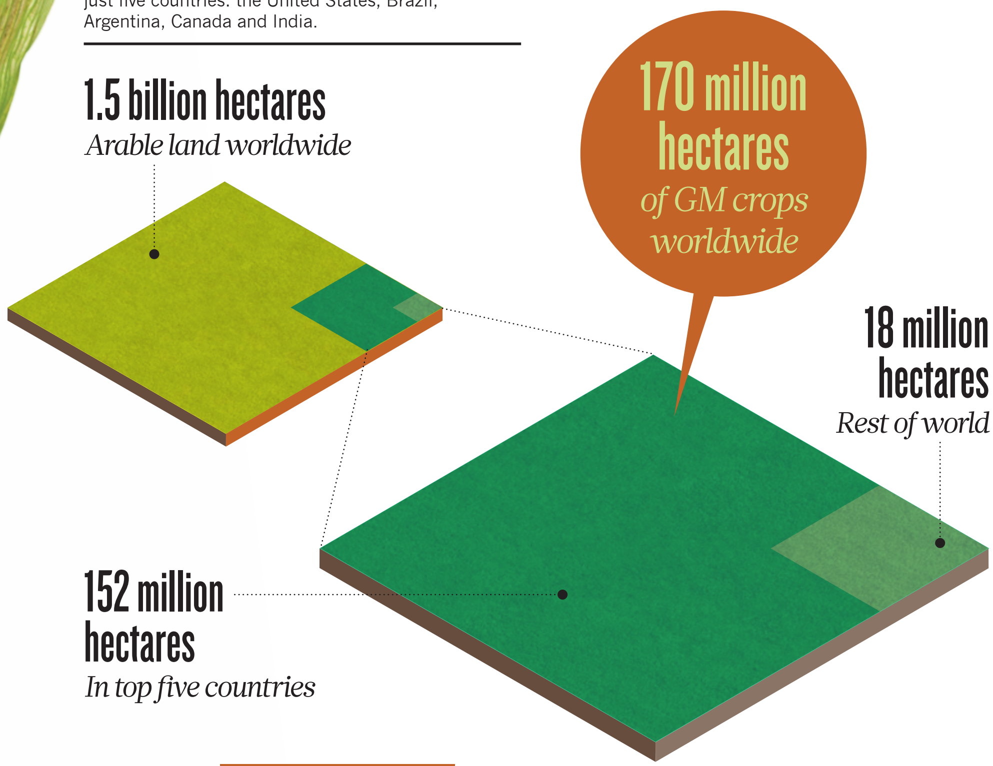
GM CROPS

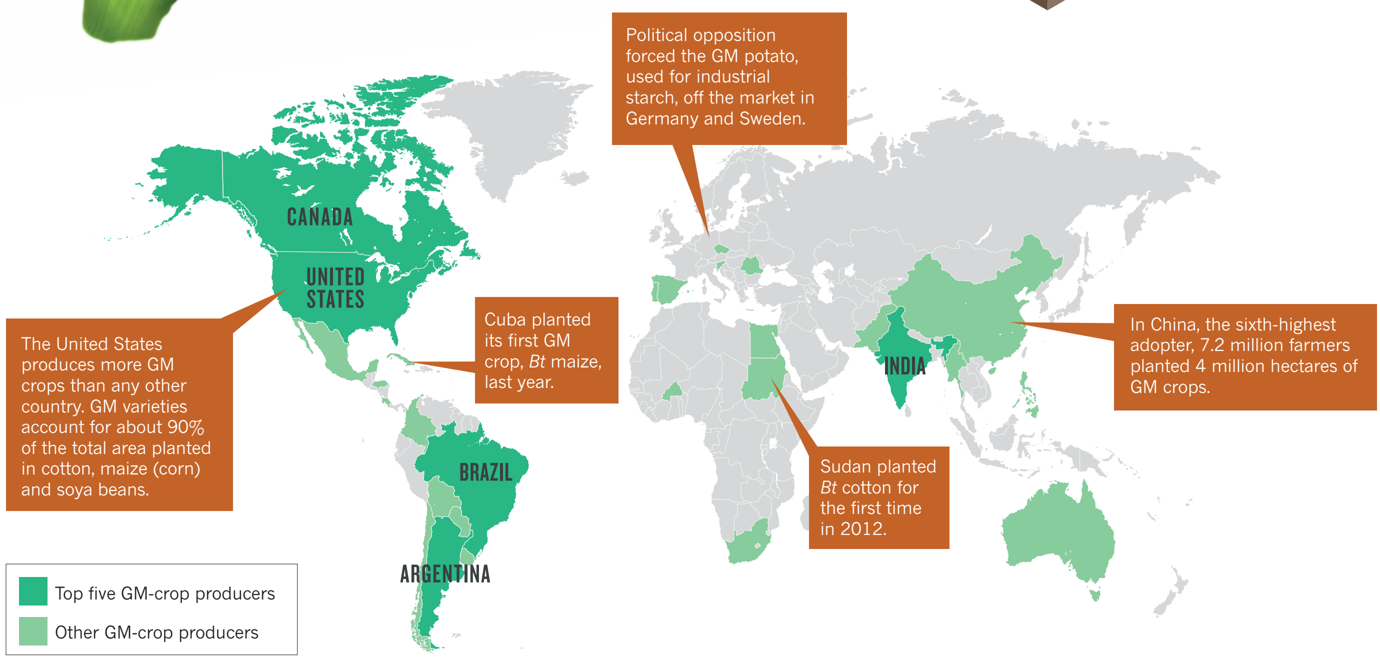
A story in numbers

In the nearly two decades since they were first commercialized, genetically modified (GM) crops have gained ground on their conventional counterparts. The vast majority are grown in five countries. Four crops feature, with two main traits: herbicide tolerance and insect resistance.

The global picture

Twenty-eight countries planted 170 million hectares of GM crops in 2012, but most crops were grown in just five countries: the United States, Brazil, Argentina, Canada and India.

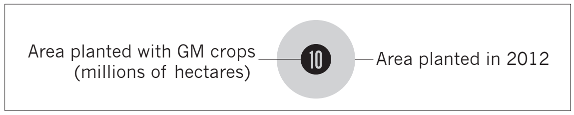


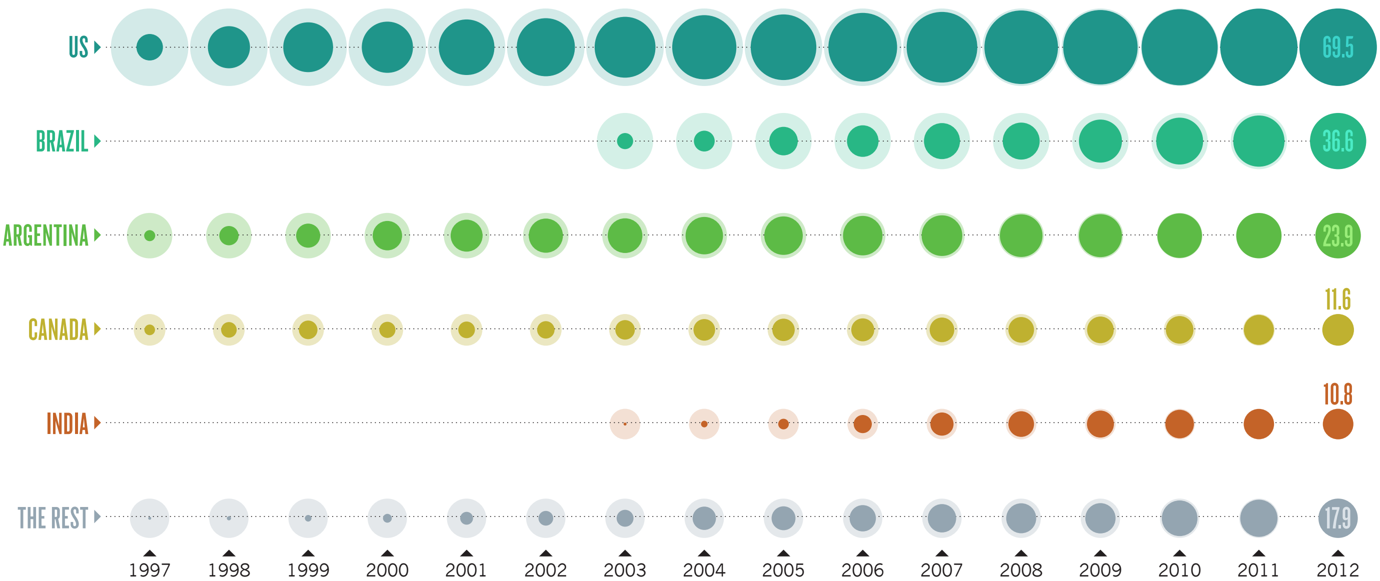


DATA PROVIDED BY THE INTERNATIONAL SERVICE FOR THE ACQUISITION OF AGRI-BIOTECH APPLICATIONS (ISAAA.ORG)

Mixed growth

Growth for many of the largest GM adopters has slowed, but Brazil is continuing to see large annual leaps with 21% (6.3 million hectares) more GM crops planted in 2012 than 2011.



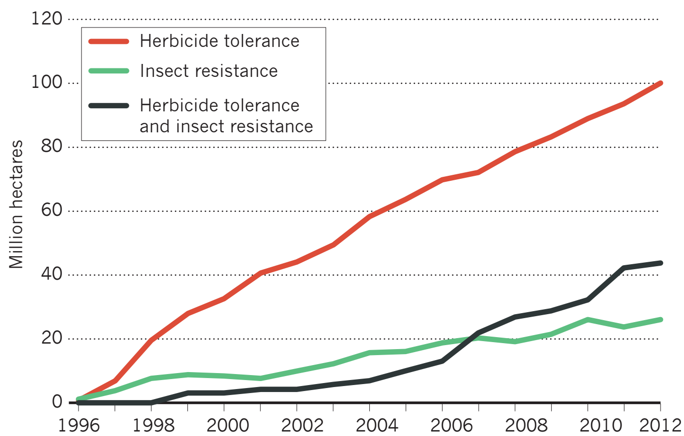


Popular crops

GM soya bean, maize (corn), cotton and canola crops accounted for nearly all GM crops grown in 2012.

Popular traits

Of some 30 traits that are currently engineered into plants for commercial use, the most popular are those that confer herbicide tolerance, insect resistance or both traits together.



Superweeds? Suicides? Stealthy genes? The true, the false and the still unknown about transgenic crops.

BY NATASHA GILBERT

A HARD LOOK AT GM CROPS

In the pitched debate over genetically modified (GM) foods and crops, it can be hard to see where scientific evidence ends and dogma and speculation begin. In the nearly 20 years since they were first commercialized, GM crop technologies have seen dramatic uptake. Advocates say that they have increased agricultural production by more than US$98 billion and saved an estimated 473 million kilograms of pesticides from being sprayed. But critics question their environmental, social and economic impacts.

Researchers, farmers, activists and GM seed companies all stridently promote their views, but the scientific data are often inconclusive or contradictory. Complicated truths have long been obscured by the fierce rhetoric. "I find it frustrating that the debate has not moved on," says Dominic Glover, an agricultural socioeconomist at Wageningen University and Research Centre in the Netherlands. "The two sides speak different languages and have different opinions on what evidence and issues matter," he says.

Here, *Nature* takes a look at three pressing questions: are GM crops fuelling the rise of herbicide-resistant 'superweeds'? Are they driving farmers in India to suicide? And are the foreign transgenes in GM crops spreading into other plants? These controversial case studies show how blame shifts, myths are spread and cultural insensitivities can inflame debate.

GM CROPS: PROMISE & REALITY
A *Nature* special issue
nature.com/gmcrops

GM CROPS HAVE BRED SUPERWEEDS: TRUE

Jay Holder, a farming consultant in Ashburn, Georgia, first noticed Palmer amaranth (*Amaranthus palmeri*) in a client's transgenic cotton fields about five years ago. Palmer amaranth is a particular pain for farmers in the southeastern United States, where it outcompetes cotton for moisture, light and soil nutrients and can quickly take over fields.

Since the late 1990s, US farmers had widely adopted GM cotton engineered to tolerate the herbicide glyphosate, which is marketed as Roundup by Monsanto in St Louis, Missouri. The herbicide–crop combination worked spectacularly well — until it didn't. In 2004, herbicide-resistant amaranth was found in one county in Georgia; by 2011, it had spread to 76. "It got to the point where some farmers were losing half their cotton fields to the weed," says Holder.

Some scientists and anti-GM groups warned that GM crops, by encouraging liberal use of glyphosate, were spurring the evolution of herbicide resistance in many weeds. Twenty-four glyphosate-resistant weed species have been identified since Roundup-tolerant crops were introduced in 1996. But herbicide resistance is a problem for farmers regardless of whether they plant GM crops. Some 64 weed species are resistant to the herbicide atrazine, for example, and no crops have been genetically modified to withstand it (see 'The rise of superweeds').

POLARIS/EYEVINE

Palmer amaranth has taken root as a herbicide-resistant 'superweed' in many US cotton fields.

Still, glyphosate-tolerant plants could be considered victims of their own success. Farmers had historically used multiple herbicides, which slowed the development of resistance. They also controlled weeds through ploughing and tilling — practices that deplete topsoil and release carbon dioxide, but do not encourage resistance. The GM crops allowed growers to rely almost entirely on glyphosate, which is less toxic than many other chemicals and kills a broad range of weeds without ploughing. Farmers planted them year after year without rotating crop types or varying chemicals to deter resistance.

This strategy was supported by claims from Monsanto that glyphosate resistance was unlikely to develop naturally in weeds when the herbicide was used properly. As late as 2004, the company was publicizing a multi-year study suggesting that rotating crops and chemicals does not help to avert resistance. When applied at Monsanto's recommended doses, glyphosate killed weeds effectively, and "we know that dead weeds will not become resistant", said Rick Cole, now Monsanto's technical lead of weed management, in a trade-journal advertisement at the time. The study, published in 2007 (ref. 1), was criticized by scientists for using plots so small that the chances of resistance developing were very low, no matter what the practice.

Glyphosate-resistant weeds have now been found in 18 countries worldwide, with significant impacts in Brazil, Australia, Argentina and Paraguay, says Ian Heap, director of the International Survey of Herbicide Resistant Weeds, based in Corvallis, Oregon. And Monsanto has changed its stance on glyphosate use, now recommending that farmers use a mix of chemical products and ploughing. But the company stops short of acknowledging a role in creating the problem. "Over-confidence in the system combined with economic drivers led to reduced diversity in herbicide use," Cole tells *Nature*.

On balance, herbicide-resistant GM crops are less damaging to the environment than conventional crops grown at industrial scale. A study by PG Economics, a consulting firm in Dorchester, UK, found that the introduction of herbicide-tolerant cotton saved 15.5 million kilograms of herbicide between 1996 and 2011, a 6.1% reduction from what would have been used on conventional cotton[2]. And GM crop technology delivered an 8.9% improvement to the environmental impact quotient — a measure that considers factors such as pesticide toxicity to wildlife — says Graham Brookes, co-director of PG Economics and a co-author of the industry-funded study, which many scientists consider to be among the field's most extensive and authoritative assessments of environmental impacts.

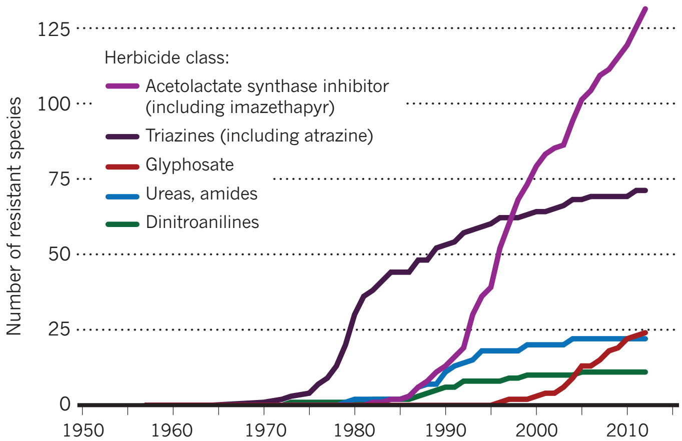
The question is how much longer those benefits will last. So far, farmers have dealt with the proliferation of resistant weeds by using more glyphosate, supplementing it with other herbicides and ploughing. A study by David Mortensen, a plant ecologist at Pennsylvania State University in University Park, predicts that total herbicide use in the United States will rise from around 1.5 kilograms per hectare in 2013 to more than 3.5 kilograms per hectare in 2025 as a direct result of GM crop use[3].

To offer farmers new weed-control strategies, Monsanto and other biotechnology companies, such as Dow AgroSciences, based in Indianapolis, Indiana, are developing new herbicide-resistant crops that work with different chemicals, which they expect to commercialize within a few years.

Mortensen says that the new technologies will lose their effectiveness as well. But abandoning chemical herbicides completely is not a viable solution, says Jonathan Gressel, a weed scientist at the Weizmann Institute of Science in Rehovot, Israel. Using chemicals to control weeds is still more efficient than ploughing and tilling the soil, and is less environmentally damaging. "When farmers start to use more sustainable farming practices together with mixtures of herbicides they will have fewer problems," he says.

THE RISE OF SUPERWEEDS

Weed species often become resistant to herbicides. Glyphosate resistance, once deemed unlikely, rose after genetically engineered crops were introduced in the mid-1990s.



SOURCE: IAN HEAP, INTERNATIONAL SURVEY OF HERBICIDE RESISTANT WEEDS WWW.WEEDSCIENCE.ORG/GRAPHS/SOAGRAPH.ASPX (2013).

GM COTTON HAS DRIVEN FARMERS TO SUICIDE: FALSE

During an interview in March, Vandana Shiva, an environmental and feminist activist from India, repeated an alarming statistic: "270,000 Indian farmers have committed suicide since Monsanto entered the Indian seed market," she said. "It's a genocide."

The claim, based on an increase in total suicide rates across the country in the late 1990s, has become an oft-repeated story of corporate exploitation since Monsanto began selling GM seed in India in 2002.

Bt cotton, which contains a gene from the bacterium *Bacillus thuringiensis* to ward off certain insects, had a rough start. Seeds initially cost five times more than local hybrid varieties, spurring local traders to sell packets containing a mix of *Bt* and conventional cotton at lower prices. The sham seeds and misinformation about how to use the product resulted in crop and financial losses. This no doubt added strain to rural farmers, who had long been under the pressures of a tight credit system that forced them to borrow from local lenders.

But, says Glover, "it is nonsense to attribute farmer suicides solely to *Bt* cotton". Although financial hardship is a driving factor in suicide among Indian farmers, there has been essentially no change in the suicide rate for farmers since the introduction of *Bt* cotton.

That was shown by researchers at the International Food Policy Research Institute in Washington DC, who scoured government data, academic articles and media reports about *Bt* cotton and suicide in India. Their findings, published in 2008 (ref. 4) and updated in 2011 (ref. 5), show that the total number of suicides per year in the Indian population rose from just under 100,000 in 1997 to more than 120,000 in 2007. But the number of suicides among farmers hovered at around 20,000 per year over the same period.

And since its rocky beginnings, *Bt* cotton has benefited farmers, says Matin Qaim, an agricultural economist at Georg August University in Göttingen, Germany, who has been studying the social and financial impacts of *Bt* cotton in India for the past 10 years. In a study of 533 cotton-farming households in central and southern India, Qaim found that yields grew by 24% per acre between 2002 and 2008, owing to reduced losses from pest attacks[6]. Farmers' profits rose by an average of 50% over the same period, owing mainly to yield gains (see 'A steady rate of tragedy'). Given the profits, Qaim says, it is not surprising that more than 90% of the cotton now grown in India is transgenic.

Glenn Stone, an environmental anthropologist at Washington University in St Louis, says that the empirical evidence for yield increases with *Bt* cotton is lacking. He has conducted original field studies[7] and analysed the research literature[8] on *Bt* cotton yields in India, and says that most peer-reviewed studies reporting yield increases with *Bt* cotton have focused on short time periods, often in the early years after the technology came online. This, he says, introduced biases: farmers who

A STEADY RATE OF TRAGEDY

Contrary to popular myth, the introduction in 2002 of genetically modified *Bt* cotton is not associated with a rise in suicide rates among Indian farmers.

SOURCE: REF. 5

adopted the technology first tended to be wealthier and more educated, and their farms were already producing higher-than-average yields of conventional cotton. They achieved high yields of *Bt* cotton partly because they lavished the expensive GM seeds with care and attention. The problem now is that there are hardly any conventional cotton farms left in India to compare GM yields and profits against, says Stone. Qaim agrees that many studies showing financial gains focus on short-term impacts, but his study, published in 2012, controlled for these biases and still found continued benefits.

Bt cotton did not cause suicide rates to spike, says Glover, but neither is it the sole reason for the yield improvements. “Blanket conclusions that the technology is a success or failure lack the right level of nuance,” he says. “It’s an evolving story in India, and we have not yet reached a definitive conclusion.”

TRANSGENES SPREAD TO WILD CROPS IN MEXICO: UNKNOWN

In 2000, some rural farmers in the mountains of Oaxaca, Mexico, wanted to gain organic certification for the maize (corn) they grew and sold in the hope of generating extra income. David Quist, then a microbial ecologist at the University of California, Berkeley, agreed to help in exchange for access to their lands for a research project. But Quist’s genetic analyses uncovered a surprise: the locally produced maize contained a segment of the DNA used to spur expression of transgenes in Monsanto’s glyphosate-tolerant and insect-resistant maize[9].

GM crops are not approved for commercial production in Mexico. So the transgenes probably came from GM crops imported from the United States for consumption and planted by local farmers who probably didn’t know that the seeds were transgenic. Quist speculated at the time that the local maize probably cross-bred with these GM varieties, thereby picking up the transgenic DNA.

When the discovery was published in *Nature*, a media and political circus descended on Oaxaca. Many vilified Monsanto for contaminating maize at its historic origin — a place where the crop was considered sacred. And Quist’s study came under fire for technical deficiencies, including problems with the methods used to detect the transgenes and the authors’ conclusion that transgenes can fragment and scatter throughout the genome[10]. *Nature* eventually withdrew support for the paper but stopped short of retracting it. “The evidence available is not sufficient to justify the publication of the original paper,” read an editorial footnote to a critique[10] of the research published in 2002.

Since then, few rigorous studies of transgene flow into Mexican maize have been published, owing mainly to a dearth of research funding, and they show mixed results. In 2003–04, Allison Snow, a plant ecologist at Ohio State University in Columbus, sampled 870 plants taken from 125 fields in Oaxaca and found no transgenic sequences in maize seeds[11].

But in 2009, a study[12] led by Elena Alvarez-Buylla, a molecular ecologist at the National Autonomous University of Mexico in Mexico City, and Alma Piñeyro-Nelson, a plant molecular geneticist now at the University of California, Berkeley, found the same transgenes as Quist in three samples taken from 23 sites in Oaxaca in 2001, and in two samples taken from those sites in 2004. In another study, Alvarez-Buylla and her co-authors found evidence of transgenes in a small percentage of seeds from 1,765 households across Mexico[13]. Other studies conducted within local communities have found transgenes more consistently, but few have been published[14].

Snow and Alvarez-Buylla agree that differences in sampling methods can lead to discrepancies in transgene detection. “We sampled different fields,” says Snow. “They found them but we didn’t.”

The scientific community remains split on whether transgenes have infiltrated maize populations in Mexico, even as the country grapples with whether to approve commercialization of *Bt* maize.

“It seems inevitable that there will be a movement of transgenes into local maize crops,” says Snow. “There is some proof that it is happening, but it is very difficult to say how common it is or what are the consequences.” Alvarez-Buylla argues that the spread of transgenes will harm the health of Mexican maize and change characteristics, such as a variety’s look and taste, that are important to rural farmers. Once the transgenes are present, it will be very difficult, if not impossible, to get rid of them, she says. Critics speculate that GM traits that accumulate in the genomes of local maize populations over time could eventually affect plant fitness by using up energy and resources or by disrupting metabolic processes, for example.

Snow says that there is no evidence so far for negative effects. And she expects that if the transgenes now in use drift to other plants, they will have neutral or beneficial effects on plant growth. In 2003, Snow and her colleagues showed that when *Bt* sunflowers (*Helianthus annuus*) were bred with their wild counterparts, transgenic offspring still required the same kind of close care as its cultivated parent but were less vulnerable to insects and produced more seeds than non-transgenic plants[15]. Few similar studies have been conducted, says Snow, because the companies that own the rights to the technology are generally unwilling to let academic researchers perform the experiments.

In Mexico, the story goes beyond potential environmental impacts. Kevin Pixley, a crop scientist and the director of the genetic resources programme at the International Maize and Wheat Improvement Centre in El Batan, Mexico, says that scientists arguing on behalf of GM technologies in the country have missed a crucial point. “Most of the scientific community doesn’t understand the depth of the emotional and cultural affiliation maize has for the Mexican population,” he says.

Tidy stories, in favour of or against GM crops, will always miss the bigger picture, which is nuanced, equivocal and undeniably messy. Transgenic crops will not solve all the agricultural challenges facing the developing or developed world, says Qaim: “It is not a silver bullet.” But vilification is not appropriate either. The truth is somewhere in the middle. ■

Natasha Gilbert *writes for* Nature *from Washington DC.*

1. Wilson, R. G. *et al. Weed Technol.* **21,** 900–909 (2007).
2. Brookes, G. & Barfoot, P. *GM Crops Food*; preprint at http://go.nature.com/q8rmke (2013).
3. Mortensen, D., Egan, J. F., Maxwell, B. D., Ryan, M. R. & Smith, R. G. *BioScience* **62,** 75–84 (2012).
4. Gruère, G. P., Mehta-Bhatt, P. & Sengupta, D. Bt *Cotton and Farmer Suicides in India.* Discussion paper 00808 (International Food Policy Research Institute, 2008).
5. Gruère, G. & Sengupta, D. *J. Dev. Stud.* **47,** 316–337 (2011).
6. Kathage, J. & Qaim, M. *Proc. Natl Acad. Sci. USA* **109,** 11652–11656 (2012).
7. Stone, G. D. *World Dev.* **39,** 387–398 (2011).
8. Stone, G. D. *Econ. Polit. Weekly* **47,** 62–70 (2012).
9. Quist, D .& Chapela I. H. *Nature* **414,** 541–543 (2001).
10. Metz, M. & Fütterer, J. *Nature* **416,** 600–601 (2002).
11. Ortiz-Garcia, S. *et al. Proc. Natl Acad. Sci. USA* **102,** 12338–12343 (2005).
12. Piñeyro-Nelson, A. *et al. Mol. Ecol.* **18,** 750–761 (2009).
13. Dyer, G. A. *et al. PLoS ONE* **4,** e5734 (2009).
14. Mercer, K. L. & Wainwright, J. D. *Agric. Ecosyst. Environ.* **123,** 109–115 (2008).
15. Snow, A. *et al. Ecol. Appl.* **13,** 279–286 (2003).

BY DANIEL CRESSEY

A NEW BREED

The next wave of genetically modified crops is making its way to market — and might just ease concerns over 'Frankenfoods'.

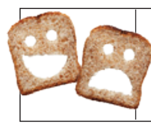
When the first genetically modified (GM) organisms were being developed for the farm, says Anastasia Bodnar, "we were promised rocket jet packs" — futuristic, ultra-nutritious crops that would bring exotic produce to the supermarket and help to feed a hungry world.

Yet so far, she says, the technology has bestowed most of its benefits on agribusiness — almost always through crops modified to withstand weed-killing chemicals or resist insect pests. This has allowed farmers to increase yields and spray less pesticide than they might have otherwise.

At best, such advances have been almost invisible to ordinary consumers, says Bodnar, a biotechnologist with Biology Fortified, a non-profit GM-organism advocacy organization in Middleton, Wisconsin. And at worst, they have helped to fuel the rage of opponents of genetic modification, who say that transgenic crops have concentrated power and profits in the hands of a few large corporations, and are a prime example of scientists meddling in nature, heedless of the dangers (see page 24).

But that could soon change, thanks to a whole new generation of GM crops now making their way from laboratory to market. Some of these crops will tackle new problems, from apples that stave off discolouration to 'Golden Rice' and bright-orange bananas fortified with nutrients to improve the diets of people in the poorest countries.

Other next-generation crops will be created



GM CROPS: PROMISE & REALITY
A *Nature* special issue
nature.com/gmcrops

IMAGE: SERGE BLOCH

using advanced genetic-manipulation techniques that allow high-precision editing of the plant's own genome. Such approaches could reduce the need to modify commercial crops with genes imported from other species — one of the practices that most disturbs critics of genetic modification. And that, in turn, could conceivably reduce the public disquiet over GM foods.

Or maybe not. Whatever promise these crops may show in the laboratory, they will still have to demonstrate their benefits in painstaking, expensive and detailed field trials; jump through multiple regulatory hoops; and reassure an often sceptical public.

SERGE BLOCH

That last part will not be easy, says Philip Bereano, who studies the political and social aspects of new technologies at the University of Washington, Seattle. He points out that the arguments over GM organisms run the gamut from concerns about safety and labelling to ethical issues with the patenting of life. "People are concerned about what they're feeding their kids," he says, "and that is not going to change."

Nevertheless, most GM-organism researchers seem convinced that the worst of the technology's problems are over, and that its future is bright. If you are looking for the jet-pack era of GM organisms, says Bodnar, "it is happening now."

The first wave of GM crops was marketed mainly to farmers, with the goal of making their jobs easier, more productive and more profitable. In 1996, for example, biotechnology firm Monsanto of St Louis, Missouri, introduced the first of its popular 'Roundup Ready' products: a soya bean equipped with a bacterial gene that allows it to tolerate a Monsanto-made glyphosphate herbicide known as Roundup. This meant that farmers could kill off the majority of weeds with one herbicide rather than several, without damaging the crop. Other GM crops soon followed, including Monsanto's *Bt* cotton: a plant modified to produce a bacterial toxin that discourages destructive bollworms and cuts down on the need for pesticides.

Farmers will continue to be a core market for the coming generation of GM organisms. At Rothamsted Research in Harpenden, UK, for example, scientists are working on GM plants that will need even less pesticide than Bt cotton, and maybe none at all. The key is an 'alarm pheromone' that some species of wild plant have evolved to mimic the chemical warning signals put out by aphids — a major crop pest in the temperate zones — when they are under attack. Putting the genes for this defence into wheat has created a crop that could trick the insects into thinking that they are in peril and drive them away. Unlike Bt cotton and other existing GM organisms, such a crop would need no insect-killing chemical for protection from pests.

Field trials are currently under way, says Maurice Moloney, director and chief executive of the Rothamsted centre. "In the greenhouse it's been very successful," he says. "If we can get it to work in the field, we'll be able to optimize it to make it a robust trait" suitable for large-scale deployment. From there, says Maloney, the team hopes to expand its efforts, searching for naturally evolved protections and deterrents in other crops, and working out how these might be enhanced or modified to fight particular pests. "For example, you could have a volatile chemical that also is a deterrent for caterpillars, stem borers and the like," says Maloney. "Potentially, if we can get this to work, the range of applications is phenomenal."

LOCAL CONCERNS

Many GM-organism researchers are pushing work on crops sometimes neglected by the big agricultural companies. In the plant biotechnology group at the Swiss Federal Institute of Technology in Zurich, for example, Herve Vanderschuren leads a team working on cassava (*Manihot esculenta*), a tropical shrub with a tuber that is a staple food in the developing world. "There is not major investment in breeding or improvement of this crop," he says.

Vanderschuren and his team are genetically engineering cassava to be resistant to two particularly damaging viruses, by starting with a variety that is naturally resistant to cassava mosaic virus, and then inserting genes that confer resistance to cassava brown streak virus. The naturally resistant strain was already tailored to local needs and markets. That kind of local adaptation is a "very important part of the research we do here", says Vanderschuren — and something that is rarely embraced by huge agribusinesses that want to sell products worldwide. Vanderschuren and his team have successfully made the plants, and are now collaborating with colleagues in Africa to arrange tests to confirm that the cassava can be grown in the field.

Much of the work on crops in developing nations focuses on nutritional enhancement. The most famous example of this effort is Golden Rice, a modified version of the staple food of half the world. Its distinct yellow hue comes from the addition of β-carotene, a precursor to vitamin A that is deficient in many East Asian diets. After much painstaking development and many objections from opponents of GM organisms — the original version of Golden Rice was announced in 2000 — the crop is currently undergoing field trials in the Philippines (see I. Potrykus *Nature* **466,** 561; 2010). It could clear the final regulatory hurdles and reach farmers by 2014.

Others have followed in its wake. James Dale, director of the Centre for Tropical Crops and Biocommodities at Queensland University of Technology in Brisbane, Australia, for example, is trying to equip bananas with resistance to Panama disease, a fungal wilt that can devastate crops, as well as increased β-carotene and a suite of other nutrients including iron. "Levels of micronutrient deficiencies are really very high" in Uganda and all across Africa, he explains, and bananas are a staple of the diet.

Field trials have already been conducted in Australia.

Although most next-generation GM organisms are aimed at farmers, some target the next step in the chain: industrial food processors. For example, Chris Dardick, a molecular plant biologist at the US Agricultural Research Service's Appalachian Fruit Research Station in Kearneysville, West Virginia, explains that it is difficult to get plums into processed foods, because removing their hard, woody cores leaves shards behind. But starting with genes from a mostly stoneless, conventionally bred plum, Dardick and his team are in the early stages of engineering a fruit with no stone at all. "Our biggest concern was how such a thing would be embraced by industry and consumers. Most of the feedback we've gotten has been quite positive," he says.

And then there are GM organisms designed to appeal directly to the final consumers. One of the first will be the Arctic Apple, which does not brown rapidly after it is cut or bitten into. This is thanks to the insertion of genes from other apple varieties that produce lower than usual levels of polyphenol oxidase, a key enzyme in the chain of biochemical events that cause browning.

"My wife and I are apple growers ourselves. We were concerned because apple consumption has been declining," says Neal Carter, president of Okanagan Specialty Fruits in Summerland, British Columbia, the developer of the Arctic Apple. Carter says that apples are losing ground in the supermarket to carrots and other fresh produce that is sold in bags, cleaned, sliced and ready to eat. Making apples that could be processed in such a way without browning could be a real boon for the industry. And if the apples are received well, says Carter, Arctic avocados, pears and even lettuce could be next.

ADVANCED TECHNIQUES

Much of the genetic-modification work so far has been achieved with relatively crude but established techniques, such as a 'gene gun' that fires gold nanopellets coated with DNA from other organisms into the cells of the target plant, which incorporate the DNA at random sites in the genome. But new tools offer unparalleled precision in editing genes. For example, enzymes called transcription activator-like effector nucleases (TALENs) and zinc-finger nucleases (ZFNs) can cut DNA at specific points chosen by the experimenter. By controlling how this break is repaired, it is possible to introduce mutations, single-nucleotide changes or even whole genes at precise sites, says Dan Voytas, who works with such techniques at the University of Minnesota in St Paul. "We can do precise insertion so we know where in the chromosome the foreign gene resides." This allows researchers to put the new gene in a spot in the genome where its expression is optimal, and reduces the risk of disrupting the plant's genome in undesirable ways. Voytas's group has already shown that tobacco plants can be modified with ZFNs to introduce herbicide resistance[1]. Other groups have added herbicide resistance to maize (corn) with ZFNs[2] or have used TALENs to snip out the gene in rice that confers susceptibility to bacterial blight[3].

But Voytas says the "real power" of these techniques lies in the ability to confer new traits by modifying native plant genes. For example, rather than engineering plants to withstand dry conditions by incorporating genes from drought-tolerant bacteria (see *Nature* **466,** 548–551; 2010), researchers could adjust the multiple native genes that help plants to survive drought. "Really, the next stage of the development of the technology is to go in and to tweak multiple genes," says Voytas.

Derek Jantz, co-founder of Precision BioSciences, a biotechnology company based in Durham, North Carolina, is also excited about working with a plant's own genes. For example, all plants have an analogue of the bacterial *EPSPS* gene that is inserted into Monsanto's Roundup Ready crops. It should be possible to create similar herbicide resistance by editing a plant's own version, rather than bringing in an external gene[4].

Like other researchers in the genetic-modification industry, Jantz declines to talk about specific research projects because of commercial confidentiality. But in general terms, he says, "what we're trying to do is take advantage of the wealth of functional genomics data that is becoming available".

A BREED APART

Some researchers are using genetic modification to accelerate conventional breeding techniques. Ralph Scorza, a plant scientist at the Appalachian Fruit Research Station, leads a team that has genetically modified plum trees. The modified trees can survive only in greenhouses. But thanks to the insertion of a gene from poplar trees, they begin to flower much earlier in their lifetimes than conventional varieties do, and then continuously thereafter. This means that researchers can breed the trees throughout the year, using selection, cross-breeding and other traditional techniques to develop traits such as disease resistance in just a few years, as opposed to the decade or more that conventional breeding might require. When the desired traits have been bred in, the transgenes that drive flowering can be bred out, leaving a modified but non-GM plant. Scorza and his colleagues are using this 'FasTrack' breeding strategy in an effort to generate resistance to the plum pox virus, and to increase the sugar content of the fruit. Researchers elsewhere are applying it to crops such as citrus.

> "The next stage of the development of the technology is to go in and to tweak multiple genes."

US regulators have already suggested that organisms modified with the newer techniques such that they contain no DNA from other species will be treated differently from conventional GM organisms. That might also alleviate public concerns. "We can overcome hopefully at least some of the opposition to the genetic modification," says Alan McHughen, a molecular geneticist at the University of California, Riverside.

Besides, notes Bondar, there may be no stopping GM organisms. She points out that genetic engineering now has a relatively low bar to entry. 'Biohackers' working with bacteria are already conducting genetic modification experiments in their garages and spare bedrooms, and there is nothing to stop them from applying their skills to plants — or animals — in the future.

"It's becoming easier all the time. I think people are hungry for this kind of thing," says Bondar. "The jet packs that everybody wanted — I think it's time for them to come out. If the marketplace isn't providing that from the top down, you may see it from the bottom up." ■ SEE NEWS FEATURE P.21

Daniel Cressey *is a reporter for* Nature *in London.*

1. Townsend, J. A. *et al. Nature* **459,** 442–446 (2009).
2. Shukla, V. K. *et al. Nature* **459,** 437–441 (2009).
3. Li, T., Liu, B., Spalding, M. H., Weeks, D. P. & Yang, B. *Nature Biotechnol.* **30,** 390–392 (2012).
4. Funke, T., Han, H., Healy-Fried, M. L., Fischer, M. & Schönbrunn, E. *Proc. Natl Acad. Sci. USA* **103,** 13010–13015 (2006).

COMMENT

AJAY VERMA/REUTERS/CORBIS

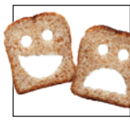
Students demonstrating against the use of GM aubergine (brinjal) in the northern Indian city of Chandigarh in 2010.

Africa and Asia need a rational debate on GM crops

Policy-makers in developing countries should not be swayed by the politicized arguments dominant in Europe, say **Christopher J. M. Whitty** and colleagues.

In Europe, scientists, politicians, industry representatives and environmentalists often present genetically modified (GM) crops either as a key part of the solution to world hunger or as a pointless but dramatic threat to health and safety. Neither position is well founded.

Recently, the often shrill debate that has unfolded in some European countries, including France and the United Kingdom, for the past 20 years has been spilling over to developing economies. The government of India, for instance, is considering banning all field trials of GM crops for the next decade — a move that could hurt large- and small-scale farmers by blocking their access to certain crop varieties that have been modified to grow better in local conditions, including types of cotton, soya bean and tomato. Meanwhile, in Kenya, where more than one-quarter of the population is malnourished, the government chose to ban the import of GM food at the end of last year but not GM crop research[1]. Like similar rulings made in Europe, such decisions seem to be based in part on emotional responses to the technology.



GM CROPS: PROMISE & REALITY
A *Nature* special issue
nature.com/gmcrops

To enable science to improve the lives of the poorest in the world, policy-makers in developing countries should resist being swayed by the politicized debate in Europe, a continent where food insecurity and malnutrition are not widely present. Instead of being either pro- or anti-GM crops, governments in developing countries should start with the specific problem at hand and assess the risks and benefits of all possible solutions — of which GM crops may be one.

Over the past 50 years, improved crop varieties have contributed almost 1% each ▶

year to the gains made in worldwide agricultural productivity[2]. In developing countries especially, new cultivars will be key in addressing the challenge of feeding rising populations in the face of climate change — along with better use of water and fertilizers, improved soil and crop management, and better storage and transport infrastructure.

Many alterations to crop varieties — to boost yields, resistance to disease and pests, nutritional value or tolerance of droughts or floods[3] — do not rely on genetic engineering. Or it may be one option among several approaches that could achieve the same result. Even in cases in which it has proved useful, genetic engineering often complements rather than supplants conventional breeding.

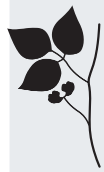
In some cases, however, it is the only viable option, for instance when there is only limited genetic variation in the trait of interest in a crop. Take the cowpea, a legume grown throughout the savannahs of Africa. Using conventional breeding, researchers have struggled for years to make cowpea resistant to a major insect pest called the Maruca pod borer (*Maruca vitrata*). The soil-borne bacterium *Bacillus thuringiensis* produces a toxin (*Bt*) that kills certain insects, including the Maruca pod borer. By crossing the *Bt* toxin gene into local cowpea varieties, researchers in Nigeria have produced resistance in 95% of plants in confined field trials (M. Ishiyaku, personal communication). In principle, *Bt* cowpea could increase yields throughout Africa by about 70% (see 'Potential life savers'). Trials on *Bt* cowpea for Maruca control are ongoing in Burkina Faso, Ghana and Nigeria, and resistant seeds will be released to farmers from 2017.

Genetic modification also offers a way to incorporate multiple traits into a plant, and to do so faster than is possible through conventional breeding.

POTENTIAL LIFE SAVERS

Genetically modified crops could transform quality of life for millions of people and boost survival rates. All three crops are in field trials.

PROVITAMIN-A-ENRICHED GOLDEN RICE
- Poor people (living on less than US$1.25 per day) eating rice each day: **400 million**
- Preschool children affected by vitamin A deficiency: **250 million**
- Deaths in children under five years that could be prevented through vitamin A provision: **>1 million**



MARUCA POD-BORER-RESISTANT COWPEA
- Consumers of cowpea in Africa: **200 million**
- Cowpea yield increase expected from pod-borer resistance: **70%**
- Reduction in insecticide spray expected with resistance: **67%**

WATER-EFFICIENT MAIZE
- Africans dependent on maize as their staple: **300 million**
- Proportion of sub-Saharan maize that suffers yield loss due to drought: **10–25%**
- Potential yield increase with drought tolerance: **20–30%**

SOURCES: GOLDEN RICE HUMANITARIAN BOARD; AFRICAN AGRICULTURAL TECHNOLOGY FOUNDATION/WEMA

Take cassava, for instance, a staple crop for millions of people in Africa. Two viral diseases — cassava mosaic disease, which stunts growth, and brown streak disease, which rots roots — affect cassava crops throughout the continent, and especially in East Africa. Varieties that are resistant to one or the other disease exist, but in many places in East Africa, both diseases are widespread. Because cassava flowers every two years, it would be enormously challenging to obtain resistance to both diseases through conventional breeding. So in Uganda and Kenya, researchers are currently investigating GM approaches.

Biofortification, whereby the nutritional value of crops is enhanced, is another area in which genetic engineering has a role. Inroads have already been made with conventional breeding methods to combat vitamin A deficiency, which can cause severe problems — for instance, by increasing the risk of childhood death from infections such as measles. An international team of researchers[4] working to improve nutrition in Mozambique and Uganda has introduced orange sweet potatoes, rich in provitamin A, into some sectors of these populations. This has translated into increased vitamin A levels in people.

In other parts of the world where sweet potatoes are not part of the staple diet, genetic modification has been used to enhance other staple crops. Producing 'golden rice', a variety genetically engineered to be rich in provitamin A, would have been impossible without using transgenic technology. Eating 150 grams of this cooked rice can provide around 60% of the Chinese recommended nutrient intake of vitamin A for 6–8-year olds[5]. Unfortunately, golden rice has not yet been approved for wide-scale use in any country, so its impact on human health has yet to be directly tested (see 'Potential life savers').

WEIGHING UP

There are good reasons for farmers in developing countries to question transgenic solutions to problems when alternatives exist. Growing non-GM crops may make better economic sense if using a GM variety would tie farmers to proprietary seeds or agrochemicals, lock them out of certain European markets, or restrict them to providing only animal feeds. The import of GM soya and maize (corn) into the European Union, for example, is currently highly regulated and limited to animal feed. Furthermore, the concern that an introduced gene will escape from one species into another with unforeseen consequences is a legitimate one, if often overstated.

Yet decision-makers in developing economies should be wary of a polarized debate that is playing out in countries where the potential benefits to society of improved crop varieties are marginal, and where people's stances towards GM foods do not necessarily reflect a considered view about the scientific technique and its alternatives.

Much of the European opposition to GM crops, although couched solely as worries about safety, also stems from concerns about the effect of large-scale farming on small-scale farmers, and the potential for biotech companies to create monopolies. In fact, people often equate all biotechnology with genetic engineering — putting the wide range of advanced non-GM techniques used to improve crops, such as tissue culture and marker-assisted breeding, into the 'unacceptable' category. These techniques can greatly assist conventional breeding efforts[6].

CARL WALSH/AURORA PHOTOS/CORBIS

In Kenya, the staple crop cassava is being genetically engineered to resist two viral diseases.

To begin with an emotional debate about GM techniques is to look down the wrong end of the telescope. Policy-makers in developing countries should instead start with the problem and make their own decisions about the balance of pros and cons of different solutions in their local context, guided by biosafety legislation.

The level of hunger and malnutrition people are currently facing in Africa and Asia, and the fact that a much higher proportion of the population in both continents depends on agriculture for their livelihoods, means that it makes little sense for decisions on GM crops to be overly influenced by European perspectives. First, by the end of the century, the United Nations estimates that less than 10% of the world's population will be living in Europe. Second, in Europe, where the benefits of better crop yields are slight, the risks (although largely theoretical, and in some cases, arguably irrational) may dominate in a risk–benefit analysis. It is worth noting that where GM technology is essential to products that Europe is short of, including some medicines, fewer concerns are expressed.

Genetic engineering is not essential, or even useful, for all crop improvements. But in some cases, it helps to improve yields and nutritional value, and reduces the risks and costs associated with the overuse of fertilizers, pesticides and water. Excluding any technology that can help people to get the food and nutrition that they need should be done only for strong, rational and locally relevant reasons. ■

Christopher J. M. Whitty *is chief scientific adviser at the UK Department for International Development (DFID), London, and professor of international health at the London School of Hygiene & Tropical Medicine, UK.* **Monty Jones** *is executive director at the Forum for Agricultural Research in Africa, Accra, Ghana.* **Alan Tollervey** *is head of agriculture research at DFID.* **Tim Wheeler** *is deputy chief scientific adviser at DFID, and professor of crop science at the University of Reading, UK.*
e-mail: c-whitty@dfid.gov.uk

1. Owino, O. *Nature* http://dx.doi.org/10.1038/nature.2012.11929 (2012).
2. Renkow, M. & Byerlee, D. *Food Policy* **35,** 391–402 (2010).
3. Varshney, R. K., Bansal, K. C., Aggarwal, P. K., Datta, S. K. & Craufurd, P. Q. *Trends Plant Sci.* **16,** 363–371 (2011).
4. Hotz, C. *et al. J. Nutr.* **142,** 1871–1880 (2012).
5. Tang, G. *et al. Am. J. Clin. Nutr.* **96,** 658–664 (2012).
6. Kijima, Y., Sserunkuuma, D. & Otsuka, K. *Dev. Econ.* **44,** 252–267 (2006).

XINHUA NEWS AGENCY/EYEVINE

Terraced fields in China, where researchers are pushing crop yields close to their biophysical limits.

An experiment for the world

China's scientists are using a variety of approaches to boost crop yields and limit environmental damage, say **Fusuo Zhang**, **Xinping Chen** and **Peter Vitousek**.

For the past two decades, commentators have hailed genetically modified (GM) crops as the magic bullet that will solve the world's food crisis. Yet obtaining the drastically bigger yields needed to feed a growing and increasingly wealthy global population — without further depleting soils, destroying natural habitats and polluting air and water — will demand an all-embracing approach.

China is taking steps towards such a strategy, and so offers an extraordinary laboratory for the rest of the world. In 2003–11, the country increased its cereal production by about 32% (more than double the world average[1]), largely by improving the performance of its least-efficient farms. Yet in the next ▶

two decades, 30–50% more food will be needed to meet China's projected demand[2]. The country has little spare land, and water shortages are reaching crisis levels in some areas. Added to this, excessive fertilizer use is a major contributor to air pollution[3] — itself a leading risk factor in hundreds of thousands of premature deaths each year. The overuse of fertilizer is also causing numerous lakes, rivers and coastal regions to become clogged with algal blooms, especially in south China.

Driven by an urgent need to both produce more food and lessen the environmental impact of agriculture — and with government money available — Chinese scientists are working out how to push crop yields close to their biophysical limits. We believe that to obtain the greatest yields for the lowest economic and environmental costs, developing and developed economies should look to China for guidance on how to integrate diverse fundamental research, including that on genetic modification, with experimental and modelling approaches.

FERTILE FARMS

In the United States, there is slightly more than half a hectare of farmed land for every person. Just one-tenth of a hectare is available for each of China's 1.3 billion people. More than 90% of the country's 230 million farms are tiny — a typical farm in the North China Plain, for instance, is around 7 metres wide and 160 metres long[4]. Owners of such farms often have other jobs in nearby towns or cities, and it rarely makes economic sense for them to invest in machinery or year-round crop or soil management when the plot is so small. Meanwhile, continual agricultural use has depleted many Chinese soils of their natural nutrient reserves, and excessive fertilization has acidified them[5]. Overall, less than 50% of the fertilizer applied to fields actually goes to the crops for which it was intended[6] — much of the rest leaches into the environment.

In the face of such challenges, agricultural scientists in China are trying to boost yields of the nation's staple grains — maize (corn), rice and wheat — by analysing fields as ecosystems. In experimental plots and in field studies on farms, they are tracking inputs and outputs such as water, nutrients, genetic material, solar and fossil energy, and muscle power from humans and animals. For instance, by chemically testing rain and irrigation water and monitoring the amount of fertilizer or manure added, researchers can estimate the amount and types of nutrient entering the system. The various inputs and outputs are optimized, to produce the greatest yield for the least resources and the lowest nutrient losses, by manipulating conditions and approaches in replicate experimental plots, and by following plots over years. For instance, in a project involving 16 universities and institutes led by the China Agricultural University in Beijing, researchers have studied the growth of wheat, maize and rice for the past 5 years in nearly 500 plots across 11 provinces. So far, they have increased both the yields and the efficiency with which the crops use nitrogen in these plots by 30–50% (unpublished data).

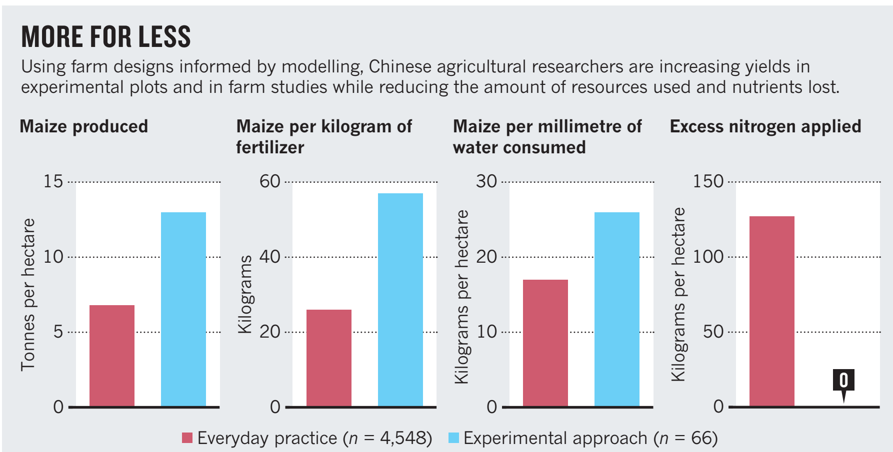
The development of new crop varieties and hybrids is one of several areas of fundamental research that feed into this approach, with transgenic technology becoming an increasingly important element in recent years. For example, by growing *Bt* cotton — the first GM crop approved for commercial use in China — farmers have increased yields since 1997 by nearly 6% and reduced the use of insecticides by around 80% (ref. 7). Although the Chinese public is wary of GM food crops, in 2008 China's central government established a 12-year research and development initiative for GM crops at a cost of 25 billion renminbi (US$3.7 billion at the time). With funding matched by the government's provincial counterparts[8], the initiative includes China's staple food crops.

Agricultural scientists are also drawing on studies of how water, nutrients and solar energy are allocated to making leaves, stems and grain; of the effect of soil structure and chemistry on roots; and of how biological, chemical and geological processes determine soil properties. Such research provides insight into the best times to add fertilizer, or the planting dates and densities that will optimize the use of water and solar energy. Recent studies on root-zone nutrient management in more than 5,000 experimental plots across 20 provinces, for instance, enabled a group led by the China Agricultural University to increase yields by 12% on average over a 7-year period. The effort also reduced fertilizer use by 24% (ref. 9).

To integrate all the relevant information

A lake in Hubei province clogged with algae, caused by excess fertilizer running off nearby fields.

REUTERS/CORBIS



Most Chinese farms are tiny, making it hard to boost yields while reducing environmental costs such as air pollution.

FARMERS: REUTERS; CYCLIST: IMAGINECHINA/CORBIS

and produce results that can be applied to China's diverse agricultural regions — from the subtropical wet south to the North China Plain and the cooler northeast — agricultural scientists are feeding information on climate, soil conditions, water supply and their variability into models. The 'optimal' genetic varieties and management approaches selected through modelling are evaluated by measuring factors such as the growth of crops and their uptake of nutrients, in experimental plots or in field studies. These measurements can then be used to improve the performance of the models, which in turn are used to further enhance yields[4] (see 'More for less').

China is not the first to use this ecosystem-modelling approach. In Europe and the United States, where the method was first developed, farmers and researchers mainly use it to fine-tune established practices. Chinese scientists, by contrast, are integrating models and experiments with nationwide monitoring networks to redesign agricultural systems on a vast scale.

DEMONSTRATING SUCCESS

By taking what they have learnt from experimental plots to real farms, Chinese scientists have already reduced fertilizer usage to economically optimal levels while maintaining yields. For instance, in a study using 49 field experiments on real farms in the North China Plain and Taihu region, researchers from various institutions were able to reduce the amount of fertilizer used by 30–60% between 2003 and 2006 without reducing yields of rice, wheat or maize[6]. (That study underpinned some of the other research on fertilizer use and efficiency already mentioned.) Other results using the modelling-ecosystem approach are also promising.

To realize its goal of a 30–50% hike in yields nationally, the Chinese government has more than tripled its investment in agricultural research since 2000, from 7 billion renminbi to 24.4 billion renminbi in 2009 — or from 0.36% to 0.66% of gross domestic product. And it has allocated 3 billion renminbi a year since 2008 to a national network of organizations for developing modern agricultural technology that, at its inception, involved 50 universities, 340 institutes, 200 companies and more than 2,000 agricultural scientists.

Even so, it remains a daunting challenge to transfer research results into farming practice across so many small farms. To support the transfer of knowledge and technology, the Chinese government is funding more than 12,000 researcher-led demonstrations of crop- and soil-management approaches throughout the country. It has established several programmes and subsidies: for instance, last year, it invested 1.5 billion renminbi to pay for soil testing to guide farmers about how much fertilizer to add to their soils and when.

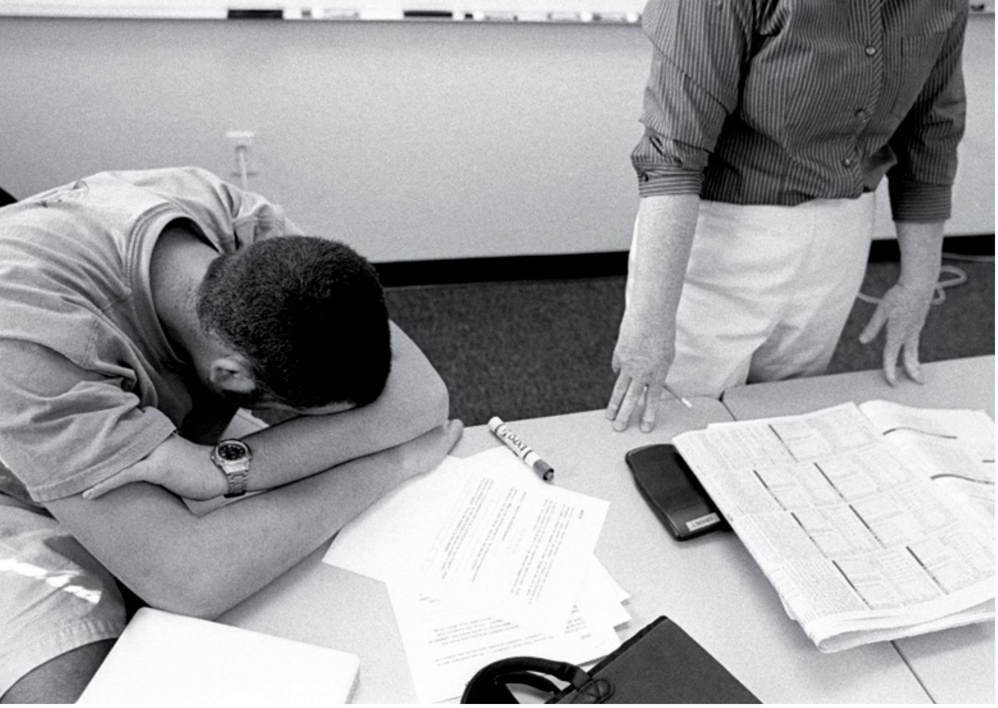
"Other developing countries should seek guidance from Chinese scientists and systems."

Chinese agricultural research must keep pace with extraordinary societal changes as well as growing demands for food, resources and environmental protection. As millions of people pour into cities such as Shanghai, Beijing and Guangzhou to find work, the agricultural heartlands of the north and south will increasingly struggle with labour shortages. Plots might become combined, with individuals managing several hectares. Also, as people continue to eat more meat and dairy items, imports of feed products will further increase; for example, in 2012, nearly 80% of the soya beans used in China were imported (58.4 million tonnes).

China's system of millions of tiny farms is unique. Yet the scope, quality and trajectory of agricultural research in China, together with its willingness and need to tackle fundamental environmental challenges, mean that other developing countries, such as India or Bangladesh, should seek guidance from Chinese scientists and systems. Farmers in Europe, North America, New Zealand and Australia can also learn from China's approach. In the face of climate change, pushing yields to the limit while sparing resources and reducing environmental consequences is a crucial goal for all. ■

Fusuo Zhang *is director, and* **Xinping Chen** *is professor at the Center for Resources, Environment and Food Security, China Agricultural University, Beijing, China.* **Peter Vitousek** *is professor in the Department of Biology, Stanford University, California, USA.*
e-mail: zhangfs@cau.edu.cn

1. *FAO Statistical Yearbook 2012* (Food and Agriculture Organization of the United Nations, 2012); available at go.nature.com/nfmwxx.
2. Zhang, F. *et al. J. Environ. Qual.* **40,** 1051–1057 (2011).
3. Liu, X. *et al. Nature* **494,** 459–463 (2013).
4. Chen, X. P. *et al. Proc. Natl Acad. Sci. USA* **108,** 6399–6404 (2011).
5. Guo, J. H. *et al. Science* **327,** 1008–1010 (2010).
6. Ju, X. T. *et al. Proc. Natl Acad. Sci. USA* **106,** 3041–3046 (2009).
7. Huang, J., Hu, R., Rozelle, S., Qiao, F. & Pray, C. E. *Aust. J. Agric. Resour. Econ.* **46,** 367–387 (2002).
8. Qiu, J. *Nature* **455,** 850–852 (2008).
9. Zhang, F. *et al. Adv. Agron.* **116,** 1–40 (2012).



ED KASHI/CORBIS

Asperger's syndrome looks set to be excluded from the world's leading psychiatric manual.

PSYCHIATRY

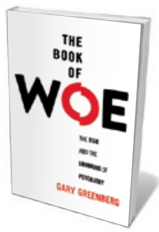
A very sad story

David Dobbs enjoys a brilliant look at the making of *DSM-5*, the new 'psychiatrists' Bible'.

When it is published this month, the fifth edition of the American Psychiatric Association's *Diagnostic and Statistical Manual*, *DSM-5*, will mark well over a century of formal guides to psychiatric diagnosis. The first two psychiatric taxonomies were produced by Emil Kraepelin in 1893 and Thomas Salmon in 1918. Kraepelin's included not only schizophrenia and what we now know as bipolar disorder, but also "masturbatory insanity" and (hopefully unrelated) "wedding night psychosis", both of which soon fell out of favour. Salmon's contained a mere 20 diagnoses. The *DSM-5*, intended to be the primary US guide to mental-health diagnoses, is expected to include some 300. These will reportedly include new entries aimed at hoarding and binge eating, and a relaxed depression diagnosis that allows therapists to classify someone grieving a loved one's death for more than two weeks as depressed rather than, well, grieving.

Each such manual, *DSM* or others, has tried to improve on its predecessor. All have failed, says psychotherapist Gary Greenberg in his entertaining, biting and essential *The Book of Woe*. But none has failed so spectacularly as the *DSM-5*.

For the first quarter of this packed but swift-reading book, Greenberg reviews how these earlier manuals — including the 1952, 1968, 1980 and 1994 editions of the *DSM* — reflected and shaped psychiatry. The history nicely sets up his main subject, which is the long, tangled efforts of the book's publisher, the American Psychiatric Association (APA), to create this fifth edition. Even as the APA started on the *DSM-5* more than a decade ago, psychiatry was suffering deep and painful divisions over issues such as overdiagnosis, overtreatment and overmedication; its problematic ties to the pharmaceuticals industry; and a shortage of demonstrable biological pathways for most diagnoses.



The Book of Woe: The DSM and the Unmaking of Psychiatry
GARY GREENBERG
Blue Rider: 2013. 426 pp. $28.95

The APA, which depends heavily on revenue from the sale of the *DSM-IV*, responded to these controversies by vigorously defending the manual — while promising to create a fifth edition that would draw on new paradigms.

From the adventure in bookmaking-by-committee that followed, Greenberg builds a splendid and horrifying read. He digs up delicious dirt; extracts from the rivalrous main players a treasure chest of kvetching, backbiting, rebuttal, regret, sibling rivalry, Oedipal undercutting and just plain pithy talk. He relates gruesome sausage-making stories about the APA's tortured attempts to refashion rusty diagnoses or forge shiny new ones. (The aetiology of that new temper-dysregulation disorder? You'll throw a fit.)

Greenberg even managed to become a tester in one of the draft manual's clinician field trials. The process proved so convoluted that he wanted to apologize to one patient for the "inadequacy, the pointlessness, the sheer idiocy of the exercise". He never got the chance: she never called again.

The *DSM*, Greenberg concludes, "dresses up symptoms as diseases that are not real and then claims to have named and described the true varieties of our suffering". Technically, the APA concurs, admitting *sotto voce* (for instance, in planning documents and public discussions for earlier versions of the *DSM*) that many psychiatric diagnoses are constructs of convenience rather than descriptions of biological ailments. This originates in an explicit decision the APA made, during the creation of *DSM-III*, to base diagnoses not on aetiology but on recognizable clusters of symptoms that seem problematic. The APA did so recognizing that this would mean stressing consistency among clinicians in recognizing symptom clusters rather than any other marker of a condition's origins.

A slippery deal, but essential. For by formalizing this scheme, psychiatry can claim medical legitimacy and accompanying insurance coverage and pay rates so that it can help people. Unfortunately, writes Greenberg, this scheme has led everyone, psychiatrists included, to talk about and treat *DSM*'s conceptual constructs as if they are biological illnesses — a habit that has bred troubles ranging from overconfidence to incestuous liaisons with Big Pharma.

Greenberg is in rough harmony with not just many clinicians but also long-time psychiatric leaders such as *DSM-IV* editor Allen Frances, whose own diatribe, *Saving Normal* (William Morrow), to be published this month, charges that the *DSM-5* will "turn overdiagnosis into hyperdiagnosis", and Thomas Insel, the psychiatrist who directs the US National Institute of Mental Health in Rockville, Maryland. Insel tells Greenberg that he hears constantly from psychiatrists who feel trapped by the *DSM*, and that perhaps it's time to "just sort of start over".

Or does the scheme work just well enough to stay afloat? As Greenberg notes, the *DSM*'s diagnoses sometimes work. For many people with Asperger's syndrome, for instance — a diagnosis introduced 19 years ago in the *DSM-IV* — the label helps them to forge a coherent identity, perhaps because it lies outside the norm from which they feel excluded. The differentiation centres rather than marginalizes them. Or, as one

NATURE.COM
For more on the evolution of the *DSM*, see.
go.nature.com/brjcau

person with the condition tells Greenberg, "It meant I'm not an asshole. I'm just wired differently."

The person who said this will probably face a dilemma on 22 May, when the *DSM-5* is published. By all reports, it will expunge Asperger's, folding it into a tightened autism spectrum disorder diagnosis. Will that person still be wired differently? Studies suggest that anywhere from one-third to three-quarters of people with Asperger's will fail to secure one of these new autism diagnoses and the accompanying health-insurance coverage and other benefits. And of those who do secure one, few are likely to find in the autism diagnosis the same satisfying fit between workable description and recognizable self-identity.

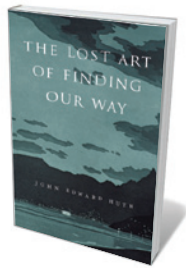
The annihilation of Asperger's suggests what may be a key part of psychiatry's tension. As Greenberg writes, the *DSM*, and psychiatry with it, increasingly "casts its subjects into dry, data-driven stories, freed from the vagaries of hope and desire, of prejudice and ignorance and fear, and anchored instead in the laws of nature". Yet when psychiatry works, it often works less at a biological than at a humanistic, narrative level, by helping the sufferer to reframe the story of his life and of his place in the world into one that includes a sense of agency, strength and social connection. This is doubtless why a combination of drugs and talk therapy generally works better than just drugs. It also helps to explain why schizophrenia, as described in Ethan Watters' *Crazy Like Us: The Globalization of the American Psyche* (Free Press, 2010) and in work by Tanya Luhrmann, is much less disabling in cultures — or even treatment regimes — that cast its eccentricities more as variations in human nature than as biological dysfunction.

For more than 100 years, psychiatry has been getting by on pseudo-scientific explanations and confident nods while it waited for the day, always just around the corner, in which it could be a strictly biological undertaking. Part of the *DSM-5*'s long delay occurred because, a decade ago, APA leaders actually thought that advances in neuroscience would allow them write a brain-based *DSM*. Yet, as former APA front liner Michael First, a psychiatrist at Columbia University in New York, confirms on Greenberg's last page, the discipline remains in its infancy.

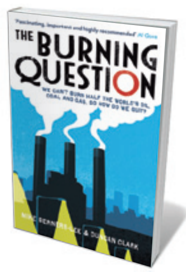
Greenberg shows us vividly that psychiatry's biggest problem may be a stubborn reluctance to admit its immaturity. And we all know how things go when you won't admit your problems. ■

David Dobbs *writes for publications including* The New York Times *and* National Geographic. *His forthcoming book,* The Orchid and the Dandelion, *focuses on the genetic and cultural roots of temperament. He blogs at Neuron Culture. e-mail: david.a.dobbs@gmail.com*

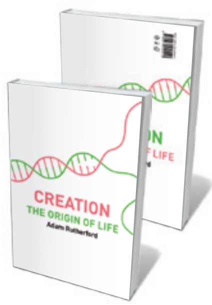
Books in brief



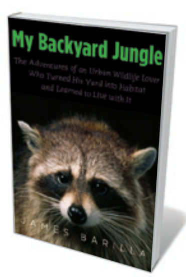
The Lost Art of Finding Our Way
John Edward Huth BELKNAP *544 pp. $35 (2013)*
Humanity's lust for exploring terra incognita shaped and tested our prodigious capacity for mental mapping. Now, with the advent of the Global Positioning System, wayfaring skills are on the wane. Physicist John Edward Huth turns explorer in this rich, wide-ranging and lucidly illustrated primer on how to find yourself in the middle of somewhere. Huth's prescription for navigating fog, darkness, open ocean, thick forests or unknown terrain rests first on harnessing compass, Sun and stars; then on the subtleties of weather forecasting and decoding markers such as the wind, waves and tides.



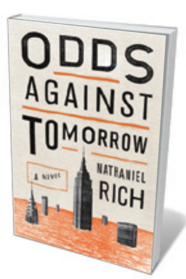
The Burning Question: We Can't Burn Half the World's Oil, Coal and Gas. So How Do We Quit?
Mike Berners-Lee and Duncan Clark PROFILE *256 pp. £9.99 (2013)*
Flabby political will and corridors of disempowerment have not dented the determination of energy writer Duncan Clark and carbon consultant Mike Berners-Lee. Arguing for a moratorium on fossil-fuel extraction, they explain why, citing the evidence on warming, the lack of an international climate-change deal, false energy 'efficiency' and the plethora of good techno-fixes. They probe the economic, social and psychological blocks to progress, and lay out a six-step solution — from pushing sustainables to capping carbon. Compelling.



Creation: The Origin of Life/The Future of Life
Adam Rutherford VIKING *272 pp. £20 (2013)*
Geneticist and *Nature* editor Adam Rutherford's two-in-one study cleverly twins the quest to understand how life emerged some 4 billion years ago with today's race to bio-engineer new life forms. In *The Origin of Life*, he marshals science history and groundbreaking recent research to build up a scenario of proto-cells spontaneously generating in the deep ocean, with a little help from RNA, lipids and mineral deposits. *The Future of Life* focuses on the potential of synthetic biology to create novel, much-needed treatments, fuels and more. Thought-provoking, and double the fun.



My Backyard Jungle: The Adventures of an Urban Wildlife Lover Who Turned His Yard into Habitat and Learned to Live with It
James Barilla YALE UNIVERSITY PRESS *376 pp. $28 (2013)*
The much-probed nexus between humans and the wild gets yet another twist in this engaging chronicle. Environmental writer James Barilla certified his garden in South Carolina as a habitat with the US National Wildlife Federation. When the experiment turned into a feral free-for-all, he sallied forth to study urban wildlife, from the garbage-scoffing macaques of New Delhi to Brazil's urban marmosets. The findings were unsettling. Ultimately, he argues, creating a "culture of coexistence" is as tough as it is necessary.



Odds Against Tomorrow: A Novel
Nathaniel Rich FARRAR, STRAUS & GIROUX *320 pp. $26 (2013)*
From Fukushima to Hurricane Sandy, catastrophes come at a hideously high price. Victims and governments feel it one way; insurers another. Nathaniel Rich lights the shadier corners of that number-crunching realm in this incisive novel. Quant Mitchell Zukor has mastered the maths of cataclysms, but his assessments are used to corrupt ends — and immersion in paper disasters fails to prepare him for the real thing. Amusing and petrifying by turns, this is near-future fiction with an edge of the real. **Barbara Kiser**

DARRYN WELCH

Michael Eklund plays a developmental biologist in the film *Errors of the Human Body*.

GENETICS

Wayward genes and grieving scientists

Alison Abbott weighs up a believable cinematic treatment of genetics research and the personalities at the bench.

Errors of the Human Body
DIRECTED BY ERON SHEEAN
Selected US cinemas: 2012.

Developmental biologist Geoff Burton is a Canadian science star, regularly churning out high-impact research papers. His charmed life implodes when his baby dies of a rare and random genetic mutation that causes uncontrolled tumour growth. When Burton sets out on a grief-stricken search for the gene, his publication rate plummets. He transfers to a leading German research institute in Dresden, only to find it fraught with tensions. By the end of this unusual arthouse medical thriller, Burton has learned that no one should hope to fully control the healing process — neither the emotional nor the physical kind.

There is a believable scientific edge to the story and details of *Errors of the Human Body*. As the characters play out their sexual and professional jealousies, they inject their mice intraperitoneally, take blood samples from the animals' tails and peer intelligently down the most modern of microscopes. With its intense themes and dispiriting blue–grey tones, the film doesn't make for easy viewing. Still, its compelling and realistic representation of the daily grind of research gives it a very special interest.

Director Eron Sheean, an Australian living in Berlin, became familiar with the scientific process during his tenure as artist-in-residence at the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden. Kai Simons, a research director at the institute, initiated the residency programme to promote understanding between intellectuals from different fields, after learning that artists had formed a key support group for a 1998 Swiss referendum that called to severely restrict genetic engineering.

At the Dresden institute, Sheean's discussions with scientists working on axolotls — a kind of salamander that can regenerate its limbs — helped him to realize the metaphorical potential of regenerative medicine in exploring human grief and healing, as well as the communication issues arising from those processes. He shot the film in early 2011, almost entirely in the institute's laboratories, animal houses and grounds, featuring scientists as extras.

Played by Canadian actor Michael Eklund, Burton arrives at the institute to find Rebekka Fiedler (Karoline Herfurth), his former postdoc and ex-lover, working as an independent group leader. She has discovered the 'Easter gene', which accelerates the speed at which axolotls can regrow detached limbs. But her ruthless and ambitious colleague Jarek Novak — played splendidly by Icelandic actor Tómas Lemarquis — is stealing her samples and experimenting secretly with the gene in mice. Novak wants the glory of translating the finding to the clinic, where it could help in human wound-healing. He has the implicit collusion of smarmy lab chief Samuel Mead, played by British comedian Rik Mayall. Trying to work out the various deceptions in his new environment, Burton steals one of Novak's mice. It has already been transfected with the Easter gene using a virus carrier so carelessly constructed that the gene is transferred to Burton when the mouse bites him.

The characters are three-dimensional. Novak lapses occasionally into Hollywood-esque mad-scientist territory, but is driven by genuine scientific curiosity and an impatience with Fiedler, who has run out of scientific ideas yet will not collaborate. Fiedler is not blameless — she is secretive about her science and manipulative in her relationships. Her Easter gene is the one that had affected Burton's baby: wound-healing and cancer are the good and evil sides of regenerative medicine. When Burton spectacularly recovers — his immune system learns to recognize his tumours as foreign — the relationship he had with his dying baby is revealed to be the most complex and heart-rending of all.

Inevitably, scientific accuracy is occasionally compromised for plot. More unfortunate, however, is that not enough scientific background is given, so many non-scientists may end up judging the film as standard science fiction. That aside, the plot is at times hard to follow, with flashbacks to Burton's pre-Dresden life unclearly signalled. The final scenes tend towards melodrama.

Screenings at the Dresden institute last October drew mixed reactions from researchers. Some were unable to suspend their disbelief. Others enjoyed the thought-provoking elements of the film and the novelty of seeing their own real science as the backdrop to a drama.

Sheean's real-life story, unlike his film, does have a happy ending. He fell in love with one of the scientists he interviewed at the institute. They married last April. ■

Alison Abbott *is* Nature*'s senior European correspondent.*

CONSERVATION

Storied rarities

Emma Marris applauds a clear-eyed look at our coy relationship with endangered animals.

When extinction looms, conservation biologists tackle the problem by studying the threatened species. What is the number of extant individuals? The diversity of the gene pool? The nature of the threats to its habitat?

Journalist Jon Mooallem takes a different tack. He pores over the stories people tell about the species. Is the polar bear, for example, a "bloodthirsty man-killer", a "delicate, drowning victim", a "cog in a Darwinian machine" or a "menacing and capable agent of its own fate"? The thesis of *Wild Ones* is that these narratives are ultimately more important for species survival than any data, management plan or science, because they determine how hard society is willing to work to keep a species going — as he puts it, "the bear is dependent on the stories we tell about it".

Mooallem talks to people on the front lines of such creatures' conservation. He profiles polar-bear activists who tell stories about the bear to encourage action on climate change; scientists who study the Lange's metalmark butterfly, now wholly reliant on a single managed habitat; and the people working to establish new migratory populations of whooping cranes by leading the birds through the sky in ultralight aircraft. Along the way, he tells other tales of the "surreal kind of performance art" that the management of wild animals has become in North America, from vaccinating ferrets to monitoring pygmy rabbits with drones and equipping red wolves with collars that can administer remote-activated sedative injections should they wander too far.

Mooallem does not come to any hard conclusions about how species preservation ought to proceed. Instead, he uses these reports to probe our emotions about the animals and our relationships with them. His conclusion is that, for the scientists and volunteers working directly with species in conservation projects, emotions are highly motivating and often complex. The people Mooallem profiles are selfless, hardworking, moral — and sometimes emotionally wounded and needy.

Wild Ones: A Sometimes Dismaying, Weirdly Reassuring Story About Looking at People Looking at Animals in America
JON MOOALLEM
Penguin: 2013.
368 pp. $27.95

A scientist going through a tough divorce throws herself into her captive-breeding work and comes to identify herself with the endangered Palos Verdes blue butterfly "as two kindred underdogs, spurned but battling their way out of a corner". A single father turns to working with whooping cranes in an effort to connect with his son — only to have this work take over much of his life. "I ran away to be a bird guy and wound up being a shitty father," he says. Of his son, he adds, "now he's in college and he doesn't even answer the phone half the time I call".

The massive impact that humans have had on Earth means that many species are hanging on by a thread. Some could be saved easily — if the political will can be summoned to protect or restore large areas of their habitat. Others will have a harder time persisting in a world increasingly dominated by human activities. Mooallem explains that for these "conservation-reliant" species, including the three showcased here, human intervention will be required indefinitely.

Conservationists lead whooping cranes' migration.

JEFFREY PHELPS/AURORA/CORBIS

For many, this is a bitter pill. We want to preserve the species, its genetic diversity and its ecological relationships. But we also want to preserve its dignity, its wildness, its indifference to us. We are chasing "an infinitely receding Eden", Mooallem writes — an aesthetic experience in which animals live in a place worthy of their nobility, with us a distant, reverent audience. We get squeamish when constant interventions to maintain species seem to diminish their wildness — for instance, by integrating them too intimately into the human landscape. Mooallem cites American crocodiles flourishing in the cooling canals of a nuclear power plant in Florida, and cranes hanging out in "a retention pond near a Walmart".

Wild Ones chronicles the emotional agony of preserving the animal, only to destroy its wildness. I found many of its stories disquieting. But I think it is important to remember that these emotions are our baggage, not the animals'. Endangered plovers, pupfish and picture-wing flies know nothing of dignity. But they can know death. ■

Emma Marris *is a writer and the author of* Rambunctious Garden: Saving Nature in a Post-Wild World.
e-mail: e.marris@gmail.com

Correspondence

Thirty years of transgenic plants

This month marks the 30th anniversary of the first successful introduction of a foreign gene into a plant (L. Herrera-Estrella *et al. Nature* **303,** 209–213; 1983). To overcome today's huge agricultural hurdles, we should move to a model that combines the best features of transgenic technology with those of organic and conventional farming.

Genetic engineering has revolutionized fundamental plant research and accelerated strategic improvements in crops. More than 170 million hectares of genetically modified crops were grown worldwide last year, to the benefit of the environment and society (see nature.com/gmcrops).

These achievements are founded on pioneering studies from 1947, when plant pathologist Armin Braun suggested that DNA from *Agrobacterium tumefaciens*, a bacterium that infects plants, could induce plant tumours. Subsequent work (1974–80) by the groups of Marc Van Montagu and Jeff Schell in Belgium, Mary-Dell Chilton in the United States and Rob Schilperoort in the Netherlands revealed that *A. tumefaciens* delivers a segment of its DNA into the plant's nuclear DNA using a plasmid-integration system — one of the earliest discoveries of a natural DNA-transfer mechanism. In May 1983, the Van Montagu and Schell lab deployed this system as a gene-expression vector, and the first transgenic plants became fact.

Wim Grunewald, Jo Bury *Flanders Institute for Biotechnology (VIB), Ghent, Belgium.*
wim.grunewald@vib.be
Dirk Inzé *VIB; and Ghent University, Ghent, Belgium.*

The high cost of overspecialization

In their entreaty to bring "all available data" back into the fold of phylogenetic systematics, Quentin Wheeler and colleagues attribute the epidemic of DNA sequence analyses to certain key advantages of DNA data (*Nature* **496,** 295–296, 2013). We suggest another, more basic explanation.

After years of training in understanding taxonomic groups and evaluating complex characters, scientists can find themselves overspecialized in a particular taxon, making them uncompetitive for employment and funding opportunities. Analysing DNA sequence data, which relies less on specialized taxonomic knowledge, does not exact such a high cost.

Xiaolei Huang, Gexia Qiao *Institute of Zoology, Chinese Academy of Sciences, Beijing, China.*
huangxl@ioz.ac.cn
Colin Favret *University of Montreal, Canada.*

Journals should be clear on misconduct

The next World Conference on Research Integrity in Montreal, Canada, on 5–8 May will make collaborators more responsible for the integrity of their research (see go.nature.com/lsd1p5). I believe that more pressure should also be brought to bear on scientific journals, which should publicly declare and reinforce their policies on fraudulent reporting of research results.

Journals were urged in 2010 to improve procedures for tackling allegations of misconduct and irresponsible research practices (www.singaporestatement.org). But progress has been unsatisfactory: some 40% of high-impact biomedical journals, for example, do not have authorship policies, let alone policies to define, prevent and punish misconduct (X. Bosch *et al. PLoS ONE* **7**, e51928; 2012).

There is little excuse for this failure to act against a common, long-standing problem. Editorial associations and publishers have established guidelines on editors' responsibilities regarding suspected or confirmed misconduct in papers (see go.nature.com/egc43n). Automatic detection of plagiarism and image manipulation is now widespread, and compulsory disclosure of financial and non-financial conflicts of interest is becoming standard practice.

Legal disputes and other complications can embroil journals that do not publicly state their policies on misconduct. Worse, those journals serve the scientific community badly.

Xavier Bosch *Department of Internal Medicine, Hospital Clinic, University of Barcelona, Spain.*
xavbosch@clinic.ub.es

Don't judge research on economics alone

Colin Macilwain argues that scientific research and development in the West should be contributing more to economic prosperity (*Nature* **495,** 143; 2013). I disagree that this is a problem in the United States.

A 2007 report from the US National Academies indicated that advances in science and technology would benefit the US economy and underpin its competitiveness in the global job market (see go.nature.com/qnir4w). Despite the effects of the global financial crisis, this outlook holds largely true.

For example, a move by ten midwestern US states towards green energy will create 85,700 jobs, produce US$41 billion in new investment and cut utility bills by $43 billion, while reducing annual carbon emissions by an amount equivalent to that from 30 coal-powered plants (see go.nature.com/e5if8v).

Research should not be judged solely by its economic benefits. Biomedical scientists, for instance, do not generally search for disease cures to get rich. It is also difficult to place monetary value on the doubling of US life expectancy from 1850 to 2008 — mainly owing to medical research and to improvements in water and sewage treatments.

Many other big problems faced by humankind, including transportation, finite natural resources, overpopulation, environmental degradation and climate change, can be solved only by science and technology, irrespective of profit motives.

Thomas E. DeCoursey *Rush University Medical Center, Chicago, Illinois, USA.*
tdecours@rush.edu

Open-access boom in developing nations

Open-access publication is not always about making publicly funded research articles freely available (*Nature* **495,** 425; 2013). Other factors could be driving the boom in open-access publishing in scientifically emerging nations.

The Directory of Open Access Journals (go.nature.com/nsrmrb) shows that the United Kingdom has 587 open-access journals, Spain has 465, Germany has 286 and France, 185. Brazil publishes 843 — the second-highest number after the United States (1,312). India is fourth (518) and Egypt is sixth (363). Romania publishes more open-access journals than Italy (264 and 256, respectively), and Turkey, Colombia and Iran each publish more than France.

Few open-access journals from the developing world are internationally recognized, however, or listed in scientific databases such as PubMed. This omission excludes the journals from impact calculations and limits the pool of international peer-reviewers, undermining the rigour of articles and their value to the public.

Despite this, the proliferation of publications in such local open-access journals can promote researchers' careers in countries in which academic evaluation depends mainly on the number, not quality, of publications.

Jagadeesh Bayry *Institut National de la Santé et de la Recherche Médicale, Paris, France.*
jagadeesh.bayry@crc.jussieu.fr

NEUROSCIENCE

Navigation with a cognitive map

Hippocampal place cells encode information about an animal's spatial world. A study now finds that these same neurons envisage a future journey moments before a rat sets off. SEE ARTICLE P.74

BRANDY SCHMIDT & A. DAVID REDISH

Navigation is a cognitive process that depends on more than a paper map; much like a modern GPS, it requires the ability to plan routes using the map. In this issue, Pfeiffer and Foster[1] (page 74) show that sequences of activity in neurons called place cells, located in the hippocampus of the rat brain, transiently predict (plan) the journey that the animal is about to take. This report provides direct evidence for the future-focused navigational activity of place cells in a realistic two-dimensional environment*.

In the 1940s, the psychologist Edward Tolman proposed that mammals (including rats and humans) have a 'cognitive map' that represents the spatial environment. He also proposed that animals could use the map to plan future trajectories[2]. In the 1970s, two neuroscientists, John O'Keefe and Lynn Nadel, suggested that the hippocampus was a key component of this cognitive map, with its place cells representing locations within the environment[3]. In the intervening years, evidence has mounted that place cells can be used cognitively — that is, they play out information about both the environment (during rest and sleep)[4–7] and potential options before starting off on a journey along a track or making a decision at a choice point in a T-shaped track[8–10].

These previous studies used one-dimensional, limited paths and so could not determine whether the hippocampus was checking specific options or actually planning future paths. Pfeiffer and Foster overcome this problem by bringing together a 40-tetrode microdrive and an elegant experimental task, which allowed them to decode sequences of two-dimensional position representations from sequences of hippocampal electrical firing. The microdrive permitted simultaneous recordings of 250 place cells, providing sufficient coverage to decode the location represented by this cell ensemble, even at short time scales (20 milliseconds). In the experimental task, rats alternately foraged for food rewards between randomly distributed locations and a stationary 'home' location that changed daily, but remained constant within each day. This combination of daily changes but consistency within a day meant that rats could learn the general task of alternately foraging and returning home, but would take novel routes that could be studied in two dimensions.

*This article and the paper under discussion[1] were published online on 17 April 2013.

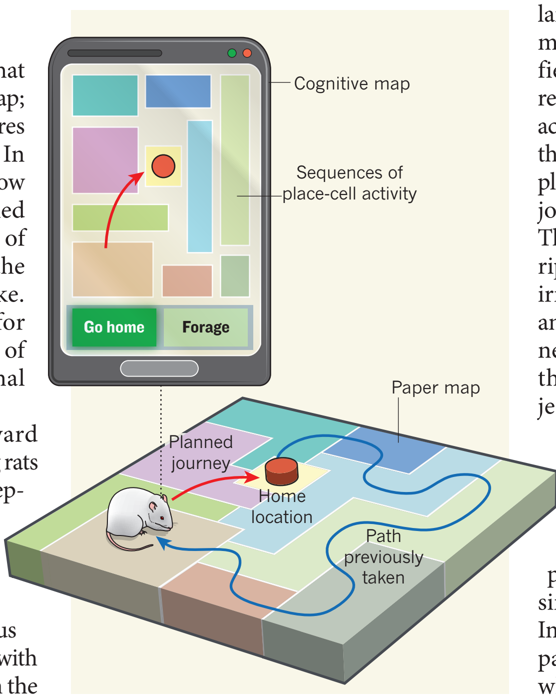


Figure 1 | What is the best way home? An ordinary paper map encodes location only, whereas a cognitive map is also involved in planning a route. Pfeiffer and Foster report that, just before a rat takes a journey, hippocampal place cells in its brain play out sequences predicting the animal's future path. This suggests that the hippocampus functions much like a GPS unit that not only shows where you are, but also how to get home.

Hippocampal place cells express most of their activity within a specific small area (called the place field) while the rat is in that area. But they also typically fire a small number of 'extra-field' spikes at other locations in the environment. In non-linguistic animals, proving that the extra-field spikes are neither part of a representation of the animal's current location nor simply random noise is an elusive proposition[11].

Pfeiffer and Foster, however, combined their large place-cell ensembles with sophisticated mathematical analyses to show that the extra-field spikes often produce a more coherent representation of the future journey than of the actual location of the rats. In the moments when the animals paused before taking a journey, the place cells fired in a sequence that predicted the journey the animal was going to take (Fig. 1). These sequences occurred during sharp-wave-ripple (SWR) events, which are well-studied irregular bursts of brief (100–200 ms), large-amplitude and high-frequency (140–200 Hz) neuronal activity in the hippocampus. And their temporal sequence represented trajectories to behaviourally relevant locations such as the next foraging location or the home base.

As an elegant control, the authors found that the representations were unrelated to the journey just completed (the past). Moreover, these trajectories were not simply straight-line paths in front of the rat. Instead, excitingly, they represented the future path taken regardless of the rat's orientation, which implies that the sequences reflected not the animal's spatial view but rather its intentions.

Future-trajectory planning was not simply a product of experience either, as it was seen even before novel traversals to the home location. As with previous examples in which untaken paths have been found to play out during waking SWR events (see ref. 7, for example), the sequences in Pfeiffer and Foster's study occurred in situations in which the map was known (the animals had a lot of experience with the environment) but the specific path to be taken was not — the home location changed every day. This is when maps are most useful, when they allow one to attach new significance to old locations[2,12].

Place-cell sequences during SWR events were originally seen during sleep and are believed to facilitate memory consolidation, which involves information transfer from the hippocampus to the cortex of the brain[4]. Indeed, disruption of SWR events during sleep impairs memory consolidation[13,14]. Increasing evidence suggests, however, that when SWR events occur during waking states they encode different information[5,15]. For instance, disruption of SWR events during wakefulness impairs only

hippocampal-dependent spatial navigation, suggesting that SWR events facilitate cognitive processes during wakefulness[16]. In their two-dimensional set-up, Pfeiffer and Foster show that the sequences during waking states reflect future plans rather than past experiences.

Functional connectivity within the hippocampal formation changes during distinct behavioural states. Whereas SWR events occur during sleep or quiet wakefulness, large-amplitude, low-frequency theta oscillations (4–12 Hz) characterize neuronal activity when an animal moves and during attentive wakefulness. Hippocampal firing during these theta states have been found to encode potential future options. For example, animals making decisions at a choice point on a T-shaped maze also show future-representing sequences, but these sequences occur during theta oscillations rather than SWR events[9,17]. A fascinating question is, what is the relationship between these two planning phenomena? Does one negate the need for the other?

It also remains unclear what triggers the hippocampal neural sequences associated with future-trajectory planning and how these sequences interact with other neural circuits. The hippocampus is only part of a complex neural network that involves several related brain structures. In humans, for example, planning processes entail an interaction of multiple structures, including prefrontal cortex[18,19]. What are these other structures doing during the planning events observed by Pfeiffer and Foster? In light of their remarkable results, researchers must now explore what processes generate these place-cell sequences, and how they are used in recalculating the journey home. ■

Brandy Schmidt *and* **A. David Redish** *are in the Department of Neuroscience, University of Minnesota, Minneapolis, Minnesota 55455, USA.*
e-mails: schmidtb@umn.edu; redish@umn.edu

1. Pfeiffer, B. E. & Foster, D. J. *Nature* **497,** 74–79 (2013).
2. Tolman, E. C. *Psychol. Rev.* **55,** 189–208 (1948).
3. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon, 1978).
4. Sutherland, G. R. & McNaughton, B. *Curr. Opin. Neurobiol.* **10,** 180–186 (2000).
5. Foster, D. J. & Wilson, M. A. *Nature* **440,** 680–683 (2006).
6. Davidson, T. J., Kloosterman, F. & Wilson, M. A. *Neuron* **63,** 497–507 (2009).
7. Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S. & Redish, A. D. *Neuron* **65,** 695–705 (2010).
8. Diba, K. & Buzsáki, G. *Nature Neurosci.* **10,** 1241–1242 (2007).
9. Johnson, A. & Redish, A. D. *J. Neurosci.* **27,** 12176–12189 (2007).
10. Singer, A. C., Carr, M. F., Karlsson, M. P. & Frank, L. M. *Neuron* **77,** 1163–1173 (2013).
11. Johnson, A. Fenton, A. A., Kentros, C. & Redish, A. D. *Trends Cogn. Sci.* **13,** 55–64 (2009).
12. Tse, D. *et al. Science* **316,** 76–82 (2007).
13. Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G. & Zugaro, M. B. *Nature Neurosci.* **12,** 1222–1223 (2009).
14. Ego-Stengel, V. & Wilson, M. A. *Hippocampus* **20,** 1–10 (2010).
15. Wikenheiser, A. M. & Redish, A. D. *Hippocampus* **23,** 22–29 (2013).
16. Jadhav, S. P., Kemere, C., German, P. W. & Frank, L. M. *Science* **336,** 1454–1458 (2012).
17. Gupta, A. S., van der Meer, M. A., Touretzky, D. S. & Redish, A. D. *Nature Neurosci.* **15,** 1032–1039 (2012).
18. Spiers, H. G. & Maguire, E. A. *Neuroimage* **31,** 1826–1840 (2006).
19. Voss, J. L. *et al. Proc. Natl Acad. Sci. USA* **108,** E402–E409 (2011).

EARTH SCIENCE

Small differences in sameness

Fresh evidence shows that the iron isotopic composition of Earth's silicate component does not, as was previously thought, reflect the formation of the planet's core at high pressure nor losses of material to space.

ALEX N. HALLIDAY

Writing in *Earth and Planetary Science Letters*, Craddock *et al.*[1] provide strong evidence that iron-isotope differences between planetary samples reflect the origins of the samples themselves rather than isotopic fractionation during planet formation. Although this is a negative result, it says a lot about planet and core formation.

Meteorites provide an invaluable archive of the circumstellar disk from which the terrestrial planets and asteroids formed. With the advent of accurate mass spectrometry and its application to meteorite samples, it was soon shown that this disk had relatively uniform isotopic compositions. For example, the uranium found on Earth has the same atomic mass as the uranium found in meteorites from the asteroid belt that lies between Mars and Jupiter, showing that the mix of isotopes from diverse primordial circumstellar-disk material was about the same. With the more recent development of a technique called multiple-collector inductively coupled plasma mass spectrometry, it has become possible to explore this 'sameness' to much higher precision and in many more elements. This has led to a search for small, mass-dependent isotopic differences that may have been imposed by effects that acted to separate (fractionate) the isotopes during planet formation. The resolution of such effects could help to confirm or refute theories about the dynamic processes that formed Earth and its metallic core[2–4].

This new area of isotope geochemistry has generated strong debate, because most isotopic differences reported so far have been small — less than about 100 parts per million per atomic mass unit (p.p.m. per AMU) — and have been obtained at the technical limits of what can be reliably resolved. Equally heated have been debates about whether systematic isotopic differences between samples can be scaled up to define planetary compositions at all. This is the focus of Craddock and colleagues' study.

Mass-dependent isotopic fractionation could in principle result from loss of planetary material to space through vaporization[2–5], or loss to a planet's core during core formation[6,7] (Fig. 1). In both cases, there could be a slight difference in terms of the ease of incorporation of a lighter isotope in one phase relative to another. These two phases could be vapour and liquid in the case of material lost to space, or silicate and metallic liquids in the case of core formation. Both of these processes are relevant to Earth's formation by mass accretion, which probably occurred at the same time as core formation by means of a series of large, stochastic, gravity-driven collisions over tens of millions of years.

Chemistry-based arguments have accumulated that Earth, and/or the various protoplanets that it incorporated during accretion, may have lost material to space from their outer silicate portions through erosion during the impacts[3]. As Earth became bigger, the gravitational energy released by accretion would have generated temperatures at which silicates and metals should have been vaporized[2]. Major growth phases through collisions are termed giant impacts, and the last such collision between Earth and a smaller planet named Theia, often referred to as 'the giant impact', led to the formation of the Moon from condensation and accretion in the resultant disk of vapour and debris[2,4]. If some of the material was lost to space rather than re-accreted to Earth and the Moon, then elements that should have only partially entered the vapour phase at these temperatures and pressures, such as lithium, silicon and iron, might show resolvable isotopic differences.

The iron isotopic composition of lunar basalt rocks has been found[5] to be on average slightly enriched in the heavier iron isotopes (about 30 p.p.m. per AMU) compared with most terrestrial mantle-derived samples, mainly basalts. The average for data from Earth is in turn

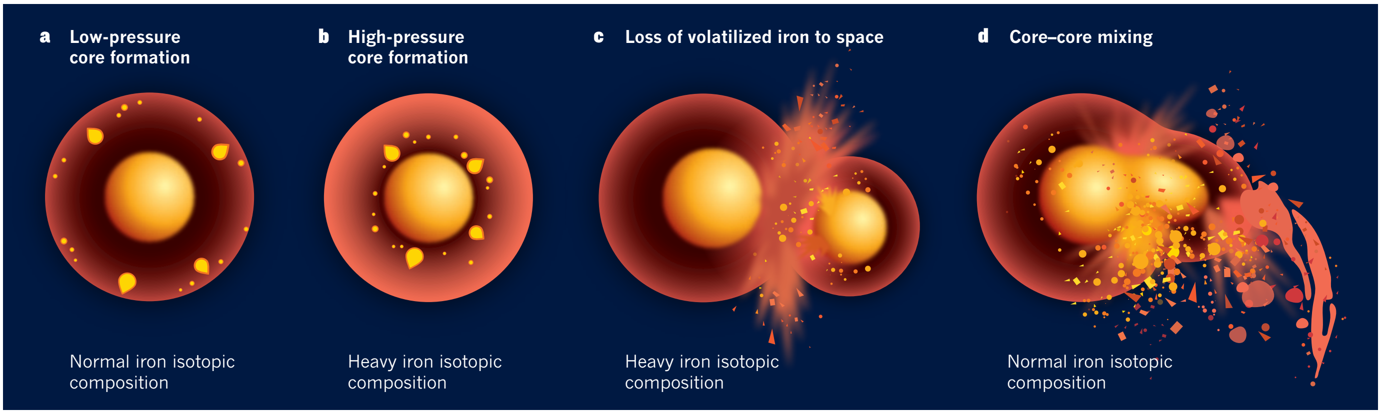


Figure 1 | Earth's formation and iron isotopic composition. Earth formed by the cumulative accretion of smaller planets and impactors. Melting from the accretion energy of these impactors would have led to segregation of dense metal (yellow) from the residual silicate of the planet (red to black, with black denoting a lower degree of melting), resulting in concomitant growth of its metallic core. The figure illustrates schematically how the iron isotopic composition of the silicate part could have been modified or left unchanged during this process, depending on the conditions of accretion and core formation. **a**, Formation of the core at low pressure is thought to leave the composition unchanged[6]. **b**, Conversely, this composition should become heavy if the core formed at high pressures[6,10]. **c**, If volatilized iron is lost to space during a collision between the proto-Earth and a small planet, then the iron isotopic composition of the residual silicate Earth should be heavy[5]. **d**, If the proto-Earth grows by repeatedly colliding with planets with low-pressure cores and the metal mixes directly with metal and silicate with silicate[13–15], the core will grow without a change in iron isotopic composition. Craddock and colleagues' results[1], coupled with earlier findings[6,8,10,11], provide evidence that **b** and **c** either resulted in no change in the iron isotopic composition, contrary to expectation, or were unimportant in the later history of Earth accretion and core formation.

slightly heavier (about 30 p.p.m. per AMU) than that for data on basalts from Mars and the asteroid Vesta. Further work confirmed that lunar basalts can indeed have a heavy iron isotopic composition, although this depends on the types of basalt analysed[8]. Furthermore, lunar basalts do not have a heavy isotopic composition for the light element lithium[9], which seems inconsistent with the idea that there were losses of lighter isotopes of iron during vaporization in the Moon-forming giant impact.

It has been argued instead that the Moon's apparent heavy iron isotopic composition might simply reflect that of the outer silicate part of Earth, which in turn was heavy because of the high pressure involved in core formation[6]. Recently, it was found that iron isotopes can become fractionated as a result of 'disproportionation' of ferrous iron into core-forming metal and oxidized ferric iron in the presence of perovskite minerals in the mantle[10]. Separation of this metal to the core is one mechanism that might explain why the silicate Earth is oxidized and why iron in terrestrial and lunar basalts is isotopically heavy, as it is for silicon[7].

More detailed studies of Earth[1,8,11], for which plentiful samples of the solid mantle are available, have raised the question of whether basalts are representative of planetary composition at all. In their favour, the mantle is compositionally heterogeneous, so individual fragments are not always representative. By contrast, basalts are derived by partial melting of large volumes of mantle and therefore provide a more effective method of averaging planetary heterogeneity. However, Craddock and colleagues' study clearly demonstrates that solid-mantle samples that have undergone melting have a lighter iron isotopic composition than basalts because of fractionation during melting. Furthermore, measurements[11,12] of chondrites, a group of primitive meteorites with a similar chemical composition to that of the Sun (if volatile elements are subtracted), show that the iron isotopic composition of the silicate Earth is like that of chondrites and so is no different from that of the Sun and the average Solar System.

These results imply that high-pressure iron-isotope fractionation, which has been demonstrated both theoretically[6] and experimentally[10], did not in fact substantially affect the silicate Earth's residual iron. The disproportionation of ferrous iron in the presence of perovskite may not have been the mechanism by which the iron in the silicate Earth became oxidized. Also, alternative models for core formation exist that do not involve segregation of metal at high pressures. For example, Earth's core and residual silicate may have grown in part in a more direct fashion by accreting smaller planetary objects that had their own low-pressure cores and by separate admixing of these objects' metal and silicate reservoirs through density differences[13]. There is supporting evidence for such core–core mixing, both theoretically[14] and in the silicate Earth's isotopic[13] and chemical[15] composition.

For some elements, such as silicon, the isotopic composition of the mantle is not fractionated greatly by melting and is heavy relative to both chondrites and samples from Mars and Vesta, with the most likely explanation being partitioning into the core[7]. The search is now on to determine which other elements have been isotopically fractionated by core formation and what this tells us about processes and conditions in the early Earth.

Craddock and colleagues' results also raise questions about the assumptions that have been made in simply comparing the iron isotopic compositions of basalts from Earth, the Moon, Mars and Vesta. It has been demonstrated over the past 10 years that the isotopic compositions of nearly all elements are the same in Earth and the Moon, leading to new models of lunar origins[4]. On this basis, the iron isotopic composition of the bulk Moon is probably also like that of chondrites, and the data for lunar basalts[5,8] may also reflect fractionation, but fractionation associated with melting on the Moon.

Recently, it has been argued that zinc isotopes in lunar samples were fractionated during the giant impact[16]. The new data from Craddock *et al.* greatly strengthen the argument that mass-dependent isotopic fractionation of elements less volatile than zinc, such as iron, magnesium and lithium, did not occur as a result of losses to space. If material was lost to space during accretion[3], it happened without isotopic fractionation of these elements, possibly because accretion was not as energetic as has been thought. Intriguingly, some of the latest simulations of the Moon-forming giant impact provide some support for this latter view[4]. ■

Alex N. Halliday *is in the Department of Earth Sciences, University of Oxford, Oxford OX1 3AN, UK.*
e-mail: alexh@earth.ox.ac.uk

1. Craddock, P. R., Warren, J. M. & Dauphas, N. *Earth Planet. Sci. Lett.* **365,** 63–76 (2013).
2. Pahlevan, K. & Stevenson, D. J. *Earth Planet. Sci. Lett.* **262,** 438–449 (2007).
3. O'Neill, H. St. C. & Palme, H. *Phil. Trans. R. Soc. Lond. A* **366,** 4205–4238 (2008).
4. Ćuk, M. & Stewart, S. T. *Science* **338,** 1047–1052 (2012).
5. Poitrasson, F., Halliday, A. N., Lee, D.-C., Levasseur, S. & Teutsch, N. *Earth Planet. Sci. Lett.* **223,** 253–266 (2004).
6. Polyakov, V. B. *Science* **323,** 912–914 (2009).
7. Armytage, R. M. G., Georg, R. B., Savage, P. S., Williams, H. M. & Halliday, A. N. *Geochim.*

Cosmochim. Acta 75, 3662–3676 (2011).
8. Weyer, S. et al. Earth Planet. Sci. Lett. 240, 251–264 (2005).
9. Magna, T., Wiechert, U. & Halliday, A. N. Earth Planet. Sci. Lett. 243, 336–353 (2006).
10. Williams, H. M., Wood, B. J., Wade, J., Frost, D. J. & Tuff, J. Earth Planet. Sci. Lett. 321–322, 54–63 (2012).
11. Schoenberg, R. & von Blanckenburg, F. Earth Planet. Sci. Lett. 252, 342–359 (2006).
12. Craddock, P. R. & Dauphas, N. Geostand. Geoanal. Res. 35, 101–123 (2011).
13. Halliday, A. N. Nature 427, 505–509 (2004).
14. Dahl, T. W. & Stevenson, D. J. Earth Planet. Sci. Lett. 295, 177–186 (2010).
15. Rubie, D. C. et al. Earth Planet. Sci. Lett. 301, 31–42 (2011).
16. Paniello, R. C., Day, J. M. D. & Moynier, F. Nature 490, 376–379 (2012).

STRUCTURAL BIOLOGY

Active arrestin proteins crystallized

Arrestin proteins regulate cellular signalling cascades initiated by ubiquitous G-protein-coupled receptors. Crystal structures reveal that two arrestins undergo similar structural changes on activation. SEE LETTERS P.137 & P.142

VALENTIN BORSHCHEVSKIY & GEORG BÜLDT

Physical abilities, cognitive skills and many other activities of humans and other organisms are generated by the coordinated actions of millions of cells. Each cell contributes to these activities by interacting with molecules outside itself. Many of these interactions are mediated through G-protein-coupled receptors (GPCRs), which in turn activate the eponymous intracellular G proteins, and so act as the starting points for numerous cellular signalling pathways. Arrestin proteins regulate the activity of these pathways by interacting with GPCRs, thus preventing G-protein-induced signalling and/or inducing additional signalling through G-protein-independent pathways. Two papers[1,2] in this issue report the first X-ray crystal structures of arrestins in their active states*.

GPCRs are members of the largest family of membrane proteins. There are more than 800 different GPCRs, most of which are activated by small molecules (agonists). One exception is rhodopsin, a well-characterized GPCR found in the light-responsive cells of the retina. Rhodopsin is activated by photons, which change the isomerization state of retinal, its covalently attached cofactor. But all GPCRs have similar interaction partners downstream: after activation of a GPCR, signalling begins when a G protein binds to the receptor, and is terminated by the binding of an arrestin molecule, which is triggered by the phosphorylation of several amino acids at the carboxy-terminal end of the receptor.

If we understand the specific interactions between biomolecules in space and time, we can picture the molecular events that underpin macroscopic biological processes. Scientists have therefore invented numerous chemical and physical methods to study these interactions, with X-ray crystallography generally having the predominant role. Rhodopsin was the first GPCR to be crystallized and to have its structure solved to high resolution[3], and the interactions of this receptor with other molecules in the visual system have been intensively studied. To establish how GPCRs interact with arrestins on the atomic scale, the method of choice would be to crystallize the two proteins together for X-ray analysis. But in most cases this is difficult. The two papers published today report impressive procedures for obtaining crystal structures of activated arrestins for which co-crystallization had failed.

Kim *et al.*[1] (page 142) describe a method for activating p44, a naturally occurring variant of visual arrestin-1 in which 35 amino-acid residues at the C terminus have been replaced by a single alanine residue. In contrast to full-length arrestin-1 (refs 4,5), which binds only to light-activated, phosphorylated rhodopsin, p44 binds to rhodopsin with high affinity regardless of whether the receptor is activated and/or phosphorylated. The same research group had previously shown[6] that opsin (retinal-free rhodopsin) behaves much like rhodopsin in the active meta II state — the intermediate that allows the enzyme rhodopsin kinase to phosphorylate the C-terminal domain of rhodopsin so that full-length arrestin-1 can bind and deactivate the receptor. In the current study, Kim and colleagues attempted to crystallize p44 in the presence of opsin. The co-crystallization failed, but the authors did obtain crystals of p44 alone, which they analysed by X-ray crystallography.

The resulting structure revealed major conformational changes in p44 when compared with unactivated full-length arrestin-1. The salient differences are a 21° twist between the amino-terminal and C-terminal domains in p44, and local changes of loop conformations and interacting hydrogen-bonding networks. The researchers attribute these conformational changes to the active form of p44, assuming that the opsin present during crystallization caused p44 to adopt this form. Whether this structure represents fully activated p44 is open to question — the protein is highly flexible, which means that crystal-packing effects might have induced the large conformational changes observed. A previously reported structure[7] of p44 crystallized in the absence of opsin revealed much smaller conformational changes than in Kim and co-workers' study.

In the second arrestin paper, Shukla *et al.*[2] (page 137) describe the crystal structure of

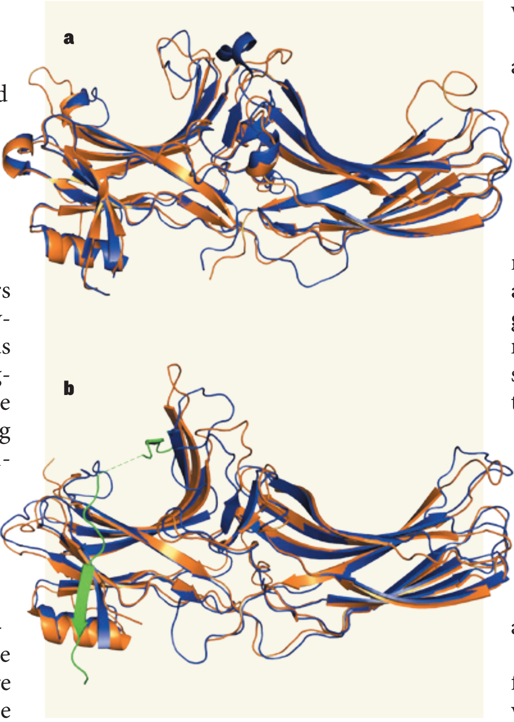


Figure 1 | Comparison of inactive and active arrestins. a, The structures of visual arrestin[5] (orange) and β-arrestin-1 (ref. 8; blue) in their inactive states are shown superimposed on each other. The N-terminal domain is on the left, and the C-terminal domain is on the right. **b,** Here, the active-state structure of visual p44 (a naturally occurring 'splice variant' of visual arrestin) reported by Kim *et al.*[1] (orange) is superimposed on the active structure of β-arrestin-1 described by Shukla and colleagues[2] (blue). The structure in green is a phosphorylated peptide that corresponds to 29 amino-acid residues of the C-terminal end of the V2 vasopressin receptor, used by Shukla *et al.* to activate β-arrestin-1. The broken line in the peptide structure denotes some amino acids that could not be traced in the X-ray crystal study. The similarity of the pairs of structures in **a** and **b** suggests that the arrestins in **b** are fully activated. The superpositions were aligned and refined with respect to the whole structures.

*This article and the papers under discussion[1,2] were published online on 21 April 2013.

non-visual β-arrestin-1 in complex with an antibody fragment (Fab30) and a phosphorylated peptide (V2Rpp) that corresponds to 29 amino-acid residues of the C-terminal end of a GPCR (the V2 vasopressin receptor). The authors found that this peptide activates β-arrestin-1, but that they could not crystallize the protein with the peptide alone — Fab30 was also needed to stabilize the activated state of β-arrestin-1 by binding to the arrestin's convex surface. Compared with the unactivated state of arrestin[8], the most remarkable change in Shukla and colleagues' structure is the 20° twist between the N-terminal and C-terminal domains, which is similar to that observed in p44 by Kim and co-workers. However, the question arises of whether the C-terminal peptide V2Rpp induces the same conformational changes in β-arrestin-1 as the complete vasopressin receptor would do. Another issue is to what degree crystal packing and Fab30 contribute to the observed conformational changes in the β-arrestin-1.

An answer comes from the superposition of the structure of the unactivated state of visual arrestin-1 with that of β-arrestin-1 (Fig. 1a), and from the superposition of p44 and β-arrestin-1 in their active states (Fig. 1b): the structures are very similar in both cases. The striking structural similarity and the presence of Fab30 in one of the structures rule out crystal-packing effects as a cause of the twist between the N-terminal and C-terminal domains. It is therefore highly probable that Kim *et al.* and Shukla and colleagues have indeed observed fully active arrestin states. In this respect, the two papers strengthen each other's results, and pave the way for further studies of the structural basis of GPCR–arrestin interactions. ■

Valentin Borshchevskiy *and* **Georg Büldt** *are at the Institute of Complex Systems (ICS-5), Research Centre Jülich, 52425 Jülich, Germany, and at the Moscow Institute of Physics and Technology, Russia.*
e-mail: g.bueldt@fz-juelich.de

1. Kim, Y. J. *et al. Nature* **497,** 142–146 (2013).
2. Shukla, A. K. *et al.* **497,** 137–141 (2013).
3. Palczewski, K. *et al. Science* **289,** 739–745 (2000).
4. Granzin, J. *et al. Nature* **391,** 918–921 (1998).
5. Hirsch, J. A., Schubert, C., Gurevich, V. V. & Sigler, P. B. *Cell* **97,** 257–269 (1999).
6. Choe, H.-W. *et al. Nature* **471,** 651–655 (2011).
7. Granzin, J. *et al. J. Mol. Biol.* **416,** 611–618 (2012).
8. Han, M., Gurevich, V. V., Vishnivetskiy, S. A., Sigler, P. B. & Schubert, C. *Structure* **9,** 869–880 (2001).

SOLID-STATE PHYSICS

Single spins in silicon see the light

The spin state of a single erbium atom in a tiny silicon transistor has been probed using laser light. The finding opens a path towards a hybrid spin–photon quantum-computing architecture. SEE LETTER P.91

CHRISTOPH D. WEIS & THOMAS SCHENKEL

The relentless drive to scale down semiconductor devices, which are at the heart of computers, tablets and video game consoles, has long been predicted to hit a brick wall when atomic dimensions are reached. However, with smaller devices also comes increased access to new functionalities, which can be based on the control of single atoms and atomic defects[1]. Access to properties other than the commonly used electric charge of electrons in a semiconductor, such as the spin of electrons and nuclei, is a possible route to improved device performance, and perhaps even to quantum computing[2]. Indeed, researchers have demonstrated[3] robust control of the spin of an electron bound to a single phosphorus atom in silicon. In this issue, Yin *et al.*[4] (page 91) describe another exciting advance in spin control — the spin-selective 'optical addressing' of a single atom in silicon.

In their study, Yin and colleagues shine light from a precisely tuned laser on a single erbium atom that has been implanted in an electrical switch (a transistor). The laser light is absorbed by the erbium atom and changes the atom's charge state in a way that depends on its spin state (Fig. 1). The transistor is operated at cryogenic temperatures in a mode in which electric transport consists of a trickle of electrons that tunnel through the device one at a time. In this single-electron transistor (SET) mode, the device can reliably detect the ionization of a single erbium atom.

Erbium is a poster child of silicon-based photonics because its dominant optical transition, which occurs at a wavelength of about 1,540 nanometres, lies right in the transmission window of the optical fibres commonly used in telecommunication. But optical transitions of erbium atoms in silicon are relatively slow, a fact that has precluded their use in the readout of spin states of single atoms through the detection of emitted photons. A key advance of the present study is the use of a SET as a sensitive electrical sensor of the charge state of an erbium atom during optical excitation. What's more, when the authors added a magnetic field to this mix, they could also use the SET to measure the spin states of the atom's electrons and nucleus. In this way, Yin *et al.* were able to observe the eight spectral lines that are associated with the 'hyperfine splitting'

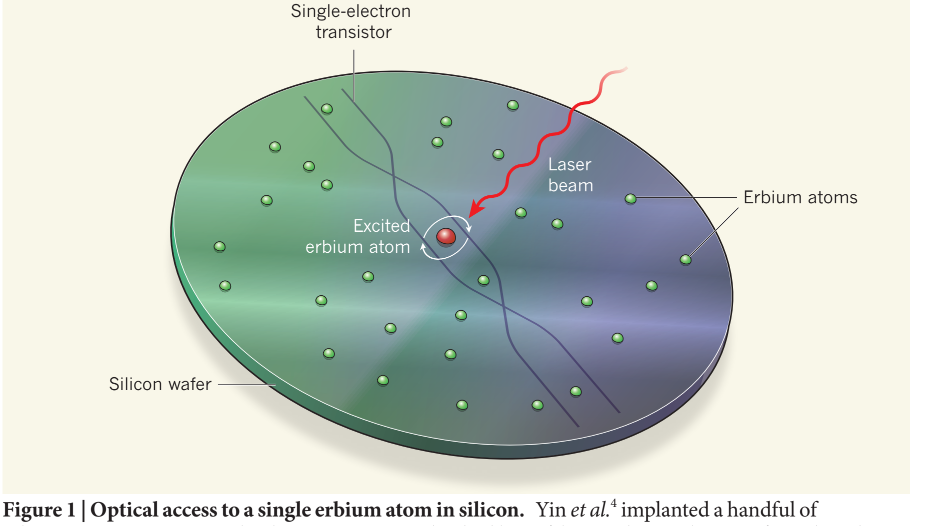


Figure 1 | Optical access to a single erbium atom in silicon. Yin *et al.*[4] implanted a handful of erbium atoms in a tiny, single-electron transistor that had been fabricated on a silicon wafer. When the wavelength of a laser beam was precisely controlled, a single erbium atom in the transistor was excited and ionized depending on the atom's spin state (white arrows); the energy levels in the atom depend on its electron and nuclear spin states, and a specific spin state can be selected by tuning the laser wavelength. The transistor senses the ionization event and thus the atom's spin state. (Figure based on an illustration by Zosia Rostomian, Lawrence Berkeley Natl Lab.)

of energy levels of one erbium atom with nuclear spin 7/2. Remarkably, the laser light used in the set-up did not cause excessive charge noise in the SET. Although the nuclear spin state was detected using multiple laser shots, single-shot detection and manipulation of single nuclear spins should become possible with further refinements of the technique.

These results are a major step towards access to both single spins and photons in a silicon-based platform — a combination that might be used to attain efficient quantum communication. The findings are also exciting because they expand the list of atoms and atomic defects for which reliable single-spin access can be obtained, beyond phosphorus in silicon[3] and nitrogen-vacancy centres in diamond[5]. And, as Yin *et al.* point out, other atomic defects can be probed with this method, too. Compared with the popular optical methods for measuring the spins of nitrogen-vacancy centres, which can be performed at room temperature, the authors' hybrid opto-electronic approach circumvents the problem of limited photon-collection efficiency that plagues purely optical methods.

It is also noteworthy that the SET used here is a FinFET[6], a three-dimensional type of transistor that is mass-produced for processors in consumer electronics. Devices based on the authors' hybrid approach might therefore serve as a platform for expanding the functionality of semiconductor systems beyond the mere shuffling of electrons to encode 'zeros' and 'ones' — the classical bits of digital information and computing. With regard to quantum computing, which requires integration of quantum memory, logic and communication modules, Yin and colleagues' work offers a promising route towards the integration of spin- and photon-based quantum bits (qubits) in silicon. This is because nuclear spins of dopant atoms such as erbium and phosphorus have extremely long 'coherence times', which are desirable for quantum memories[7]. In addition, quantum information could be transferred between electron and nuclear spins[8] and possibly also encoded in single photons emitted by an erbium atom.

To integrate erbium-based qubits into a photonic network, the issue of efficient photon collection will also have to be addressed, perhaps by placing the qubits in optical cavities (arrangements of highly reflective mirrors that trap light). And although many challenges remain, integrating these elements in silicon has the tantalizing potential to achieve distributed quantum-computing architectures (see, for example, ref. 9) at the exact wavelength that is used for classical telecommunication. ■

Christoph D. Weis *and* **Thomas Schenkel** *are in the Accelerator and Fusion Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.*
e-mail: t_schenkel@lbl.gov

1. Koenraad, P. M. & Flatté, M. E. *Nature Mater.* **10,** 91–100 (2011).
2. Awschalom, D. D., Bassett, L. C., Dzurak, A. S., Hu, E. L. & Petta, J. R. *Science* **339,** 1174–1179 (2013).
3. Pla, J. J. *et al. Nature* **489,** 541–545 (2012).
4. Yin, C. *et al. Nature* **497,** 91–94 (2013).
5. Gaebel, T. *et al. Nature Phys.* **2,** 408–413 (2006).
6. Hisamoto, D. *et al. IEEE Trans. Electr. Devices* **47,** 2320–2325 (2000).
7. Steger, M. *et al. Science* **336,** 1280–1283 (2012).
8. Morton, J. J. L. *et al. Nature* **455,** 1085–1088 (2008).
9. Van Meter, R., Ladd, T. D., Fowler, A. G. & Yamamoto, Y. *Int. J. Quantum Inform.* **08,** 295–323 (2010).

OPTICAL DEVICES

Seeing the world through an insect's eyes

An elegant combination of electronics and elastic materials has been used to construct a small visual sensor that closely resembles an insect's eye. The device paves the way for autonomous navigation of tiny aerial vehicles. SEE LETTER P.95

ALEXANDER BORST & JOHANNES PLETT

Flies are usually treated with disdain. Most commonly associated with spreading disease, they are at best considered simply annoying. Conversely, and far less appreciated, flies have also inspired mankind for centuries. An early report along these lines dates back to the seventeenth century, when the young René Descartes, while lying sick in bed, observed a fly walking along the ceiling of his room. Thinking about how he could describe the path of the fly in quantitative terms, he came up with what has become known as Cartesian coordinates, which allow algebra to be applied to geometry, and the importance of which can hardly be overestimated. The most recent example of such insect-inspired research is described by Rogers and colleagues (Song *et al.*[1]) on page 95 of this issue — the authors have transferred the design of an insect's compound eye to a digital camera.

In almost all cameras used today, the light reflected from an object in the environment is collected by a single lens and projected onto a layer of light-sensitive material in such a way that a sharp image is formed. Our eyes, as well as all other vertebrate eyes, also use this principle of image formation. The concept has the clear advantage of optimal usage of photons, guaranteeing maximum light sensitivity. Furthermore, it provides high spatial resolution, which is limited only by the density of photoreceptors in the focal plane of the lens.

Nevertheless, most living organisms use compound, or faceted, eyes instead of lensed eyes to see the world. Faceted eyes have very different optics and are composed of many hundreds or thousands of optical units (facets)[2]. In the case of the 'apposition' eye of daylight insects, each facet is optically isolated from its neighbour and equipped with its own lens and set of photoreceptors. Because each facet accepts photons from only a small angle in space, the light sensitivity of apposition eyes is rather low and the spatial resolution is limited by the number of facets that can be packed on to the small head of the insect. However, apposition eyes provide their bearer with a panoramic view of the world as well as with an infinite depth of field, without the need to adjust the focal length of the individual lenses.

UNIV. ILLINOIS & THE BECKMAN INSTITUTE

Figure 1 | Insect-inspired visual sensor. Song *et al.*[1] have designed a visual sensor that resembles, both functionally and structurally, the apposition eye of a daylight insect.

Song and colleagues now report the successful engineering of a digital camera that mimics the insect apposition eye in almost every aspect (Fig. 1). To achieve this end, the authors combined an array of elastic microlenses and a deformable array of photodetectors into a two-layer design, and transformed both layers from a planar geometry into a hemispheric shape (see Fig. 1 of the paper[1]).

The key to the success of this procedure lies in maintaining correct alignment between both sheets so as not to introduce unwanted optical artefacts. Song *et al.* attained this by

rigidly joining the two layers only at the precise locations where the microlenses overlie the photodetectors, while permitting the layers to deform independently elsewhere. The use of a turret-like, domed structure for each microlens effectively decoupled the microlenses from the mechanical stress caused by bending. Furthermore, the authors used deformable, serpentine conductor wires as a flexible electrical interconnect between photodetectors. The result is a small, artificial faceted eye with a near-hemispheric field of view, without off-axis aberration and with an almost infinite depth of field.

Given their almost complete coverage of visual space, faceted eyes are ideal for calculating the apparent motion of an object generated by its motion relative to the observer (optical flow)[3]. With regard to potential applications, the camera proposed by Song *et al.* might constitute an optimal front-end visual sensor for tiny aircraft called micro aerial vehicles (MAVs)[4]. Although, so far, most cameras on board MAVs simply use fisheye lenses to produce a wide-angle field of view[5], Song and colleagues' camera would provide all the advantages of an apposition eye. Using it to compute a MAV's self-motion could on the one hand facilitate motion stabilization in space while on the other enabling spatial navigation[6].

As with any development, there is always room for improvement. The camera's low light sensitivity, which is inherent in apposition eyes, could be ameliorated by placing more than one photodetector beneath each microlens and combining the output of photodetectors in neighbouring facets looking at the same point in space. In fact, flies use this principle of 'neural superposition' to increase the amount of light detected by the eye by a factor of seven, thereby achieving significantly higher light sensitivities[7].

Such resolvable issues aside, the system proposed by the authors could prove a stepping stone towards autonomous navigation of MAVs in their manifold possible uses. One major application is disaster relief. Picture the following: a palm-sized MAV uses an artificial faceted eye to navigate autonomously through a collapsed building while other sensors on board scan the environment for smoke, radioactivity or even people trapped beneath rubble and debris. Although these MAVs do not exist yet, thanks to devices such as that reported by Song *et al.*, they should come within reach in the foreseeable future. ■

Alexander Borst *and* **Johannes Plett** *are at the Max-Planck-Institute of Neurobiology, 82152 Martinsried, Germany.*
e-mail: aborst@neuro.mpg.de

1. Song, Y. M. *et al. Nature* **497,** 95–99 (2013).
2. Land, M. F. & Fernald, R. D. *Annu. Rev. Neurosci.* **15,** 1–29 (1992).
3. Koenderinck, J. J. & van Doorn, A. J. *Biol. Cybern.* **56,** 247–254 (1987).
4. Floreano, D., Zufferey, J.-C., Srinivasan, M. V. & Ellington, C. (eds) *Flying Insects and Robots* (Springer, 2010).
5. Plett, J., Bahl, A., Buss, M., Kühnlenz, K. & Borst, A. *Biol. Cybern.* **106,** 51–63 (2012).
6. Srinivasan, M. V. & Zhang, S. *Annu. Rev. Neurosci.* **27,** 679–696 (2004).
7. Kirschfeld, K. *Exp. Brain Res.* **3,** 248–270 (1967).

MOLECULAR BIOLOGY

The ends justify the means

A genomic analysis of yeast reveals that individual genes produce a rich complexity of RNA molecules with differing start and end sequences. The variation in these transcripts reflects the diversity of gene-regulation mechanisms. **SEE LETTER P.127**

B. FRANKLIN PUGH

It seemed simple enough, the idea that one gene encodes one RNA transcript that is translated into one protein. But over the past 50 years, molecular biology has proven to be more complex than this 'central dogma', first proposed by Francis Crick in 1958[1]. On page 127 of this issue, Pelechano *et al.*[2] take transcript complexity to new ends, reporting nearly 2 million different RNA transcripts for a yeast genome that contains roughly 6,000 protein-encoding genes*.

The RNA molecules studied by the authors are called transcript-RNA isoforms, or TIFs. These are RNAs that traverse the same region of a genome, but have differing start (5′) and end (3′) sequences. Different TIFs have the potential to alter the coding and regulatory capacity of RNA[3,4], and so the diversity of TIFs identified is intriguing. However, the plethora of isoforms may distil down to a relatively small number of functionally distinct RNAs for each gene (Fig. 1). For example, the TIFs found by the authors often have ends just a few nucleotides apart, and can be clustered into about 370,000 major TIFs (mTIFs). TIFs belonging to an mTIF probably arise from imprecise initiation of transcription. Moreover, isoforms can be quite rare compared with the predominant TIFs, which raises questions about the importance of low-abundance TIFs. The advance presented by Pelechano and colleagues' study is the comprehensiveness and resolution of TIF identification, and the complexity of transcript diversity, that can be revealed by sequencing both the 5′ and 3′ ends of the same RNA molecule.

When piecing together the full complement of RNA transcripts in a cell from sequencing data, one may be erroneously led into thinking that two overlapping transcripts represent a single longer transcript, such that distinct transcripts might go unnoticed. This problem is exacerbated by the fact that some sequencing methods fall short of reading both ends of the same RNA molecule, owing to very long transcript lengths and short-read technology. This problem was solved by methods that generate paired-end ditags (PETs)[5], in which the 5′ and 3′ ends of an RNA molecule are converted to DNA and stitched together; sequencing across this junction allows simultaneous identification of both ends of the RNA. Pelechano *et al.* present a refined version of this method, called TIF-Seq, which they use in conjunction with deep sequencing, such that each RNA sequence is detected multiple times in the data set.

A typical protein-coding messenger RNA (mRNA) molecule has three main parts: a 5′ untranslated region (UTR), a coding region (also called an open reading frame, or ORF) and a 3′ UTR. Pelechano and colleagues' analysis shows that all of these regions can be varied in a cell's TIF repertoire, and suggests how this might regulate cell function. For example, the authors identify more than 200 genes with mTIFs that start or end within the coding region, thereby resulting in truncated proteins with potentially altered function. UTRs often contain regulatory elements that alter mRNA stability and protein-translation efficiency, and the authors also find that these regulatory elements occur more often where the location of 5′ and 3′ ends is most variable. Thus, by virtue of where the enzyme RNA polymerase II, which transcribes DNA to RNA, starts or stops, the resulting RNA may have a vastly different rate of turnover or translation.

The authors also report more than 700 examples of differential UTR usage that is related to the presence of an upstream ORF (uORF). uORFs are thought to be too short to code for a protein, but they may regulate the translation frequency of the downstream ORF, thereby affecting the amount of protein produced. However, the TIF-Seq assay revealed that about half of the annotated uORFs are actually transcribed independently of the downstream ORF, suggesting that they have

*This article and the paper under discussion[2] were published online on 24 April 2013.

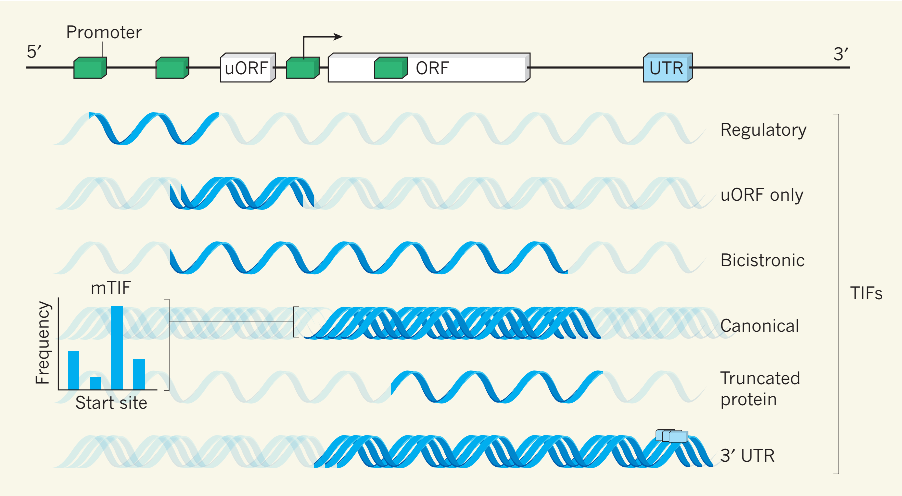


Figure 1 | Diversity of transcript isoforms. Pelechano *et al.*[2] reveal that a plethora of different RNA molecules, called transcript-RNA isoforms (TIFs), can form from transcription of sequences in and around a particular gene (top panel). TIFs include promoter-spanning regulatory transcripts, transcribed upstream open reading frames (uORFs) that may or may not be independent of a downstream protein-coding ORF, and RNA regulatory elements with untranslated regions (UTRs). Distinct promoters may result in TIFs that differ greatly in where they start, whereas imprecise initiation of transcription may produce start-site variations of only a few nucleotides. These latter TIFs, which will occur at varying frequency, can be clustered together as an mTIF.

an independent function. Reciprocally, many examples were found in which two adjacent ORFs that were thought to be transcribed independently were actually encoded on the same transcript.

Some of the identified TIFs might have no protein-coding function, such as those that run across the start site of a neighbouring gene. It is conceivable that simply the act of transcription can regulate the expression of a nearby gene, by altering the local structure of chromatin (the complex of DNA and associated proteins that make up chromosomes) and thereby changing the accessibility of DNA in the region. In this scenario, the region that RNA polymerase II transcribes will be important, but the resulting TIF will be an irrelevant by-product.

So what is the origin of TIF diversity? Pelechano and colleagues focused on the possible sources of different 5′ ends, although their results indicate that there is actually even more variation in 3′ ends. One source of 5′ variation will be at the initiation of transcription. Transcription enzymes are guided to the DNA by sequence elements, such as the TATA box, in promoter regions in DNA. In yeast, RNA polymerase II is thought to scan downstream from these sites for a certain distance before initiating transcription. Initiation might be more probable at DNA sequences at which the enzyme dwells for longer, and this might be driven directly by the underlying sequence and/or by impeding chromatin structures[6]. Variability in 5′ ends might also arise through the influence of multiple promoters on a given gene. Alternatively, weak or defective promoters might produce a small number of RNA molecules that are measurable but not meaningful.

Although TIF-Seq will help us to understand these processes by providing robust estimates of transcript diversity, it is not without its limitations. First, bona fide TIFs that are not post-transcriptionally modified by the cell (by 5′ capping and 3′ polyadenylation) will not be detected. Second, the reverse-transcriptase enzyme used to convert the RNA to DNA in the TIF-Seq assay must read the entire length of the RNA, which is problematic for long RNA molecules. TIF-Seq may therefore be less sensitive in human systems, which have long RNAs. Third, because the method detects only RNA ends, it does not detect whether isoforms have had a portion of their interior spliced out, as is common in multicellular organisms.

Despite these caveats, there is no doubt that the diversity of isoforms per cell identified by Pelechano *et al.* is dramatic. The findings help us not only to understand the origin of these transcripts, but also to explain why cell populations can never be homogeneous, even if they are derived from a single starting cell. Such intrinsic heterogeneity has broad implications, from providing opportunities for adaptation during evolution to explaining in part why it is difficult to kill all cancer cells in a tumour. ■

B. Franklin Pugh *is in the Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16803, USA.*
e-mail: bfp2@psu.edu

1. Crick, F. H. *Symp. Soc. Exp. Biol.* **12,** 138–163 (1958).
2. Pelechano, V., Wei, W. & Steinmetz, L. M. *Nature* **497,** 127–131 (2013).
3. Miura, F. *et al. Proc. Natl Acad. Sci. USA* **103,** 17846–17851 (2006).
4. Nagalakshmi, U. *et al. Science* **320,** 1344–1349 (2008).
5. Ruan, X. & Ruan, Y. *Methods Mol. Biol.* **809,** 535–562 (2012).
6. Rhee, H. S. & Pugh, B. F. *Nature* **483,** 295–301 (2012).

BIOCHEMISTRY

Oxidation controls the DUB step

Reversible oxidation of amino-acid residues can directly regulate the activity of cellular enzymes. This principle has now been extended to deubiquitinating enzymes, with implications for cell signalling and protein turnover.

MICHAEL J. CLAGUE

The oxidation of proteins is widely viewed as a signature of irreversible damage, resulting from ageing or chronic conditions[1], and cellular-degradation pathways operate to remove such proteins. However, some cysteine amino-acid residues can be reversibly oxidized to alter a protein's activity in response to prevailing conditions, in a manner akin to protein regulation by phosphorylation. Writing in *Nature Communications*, Kulathu *et al.*[2] and Lee *et al.*[3] show that cysteine oxidation occurs at the active site of deubiquitinases — a large family of enzymes that remove ubiquitin groups from proteins.

The reversible oxidation of cysteine residues is mediated by reactive oxygen species (ROS) — chemically reactive molecules that are formed during normal cellular metabolism and that are involved in many signalling pathways. ROS-mediated oxidation requires deprotonation (loss of a hydrogen ion) of the cysteine residue, and so those residues with a lower acid-dissociation constant will be especially susceptible. This condition is commonly satisfied by cysteines in the active site of hydrolytic enzymes (those that break chemical bonds in compounds such as proteins or nucleic acids). The most notable examples

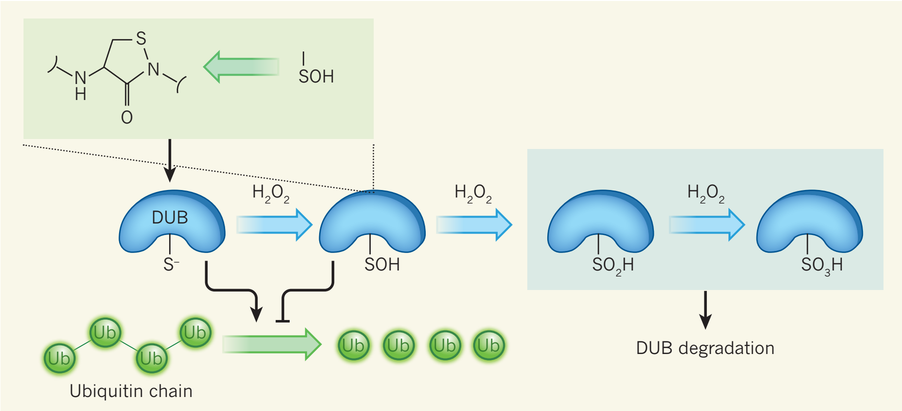


Figure 1 | Redox regulation of deubiquitinating enzymes (DUBs). The active site of many DUBs contains a cysteine amino-acid residue. Deprotonation of this residue results in the creation of a nucleophile (S^-), which allows the enzymes to attack and disassemble ubiquitin chains. However, Kulathu *et al.*[2] and Lee *et al.*[3] show that cysteine deprotonation also renders DUBs vulnerable to oxidation by reactive oxygen species, such as H_2O_2, creating a sulphenic acid group (SOH) at the cysteine residue. This oxidation inhibits the activity of the enzyme, but is reversible in the presence of a reducing agent, as long as further oxidation to sulphinic acid (SO_2H) or sulphonic acid (SO_3H) is avoided. Such oxidation would probably lead to degradation of the enzyme. Lee *et al.* show that one mechanism by which further oxidation is prevented is reaction of the SOH group with a nitrogen atom in a neighbouring residue to generate a reversible sulphenylamide species (green inset).

of enzymes affected by such oxidation have been the tyrosine phosphatases, several of which are directly inhibited by ROS[4]. Now, deubiquitinases are shown to be similarly susceptible.

Ubiquitin is a small polypeptide molecule that covalently attaches to proteins either singly or in the form of polymeric chains. Deubiquitinase enzymes (DUBs) disassemble ubiquitin chains and strip them from their substrate proteins[5]. Through this activity, they can rescue proteins from ubiquitin-dependent degradation pathways or contribute to the dynamics of ubiquitin-mediated signalling. Humans have around 80 DUBs, divided into five subfamilies, of which four comprise cysteine proteases (the USP, UCH, OTU and Josephin subfamilies) and one contains a metalloprotease (the JAMM subfamily). Although the overall architecture of the catalytic domains of the four classes of cysteine-protease DUB are highly divergent, most of the enzymes share a triad of amino-acid residues that adopts a structurally conserved disposition at the enzyme's active site. Crucially, one member of this triad is a histidine residue that deprotonates the active-site cysteine, thereby creating a nucleophilic (electron-donating) site that facilitates attack on the isopeptide bond of the substrate.

Kulathu *et al.* and Lee *et al.* used *in vitro* assays to assess how physiologically relevant concentrations of hydrogen peroxide (H_2O_2) — a source of ROS — affect the activity of purified DUB enzymes. They observed widespread inhibition of the enzymes, but show that this can be readily reversed by an excess of a reducing agent such as dithiothreitol. The authors provide multiple examples of this behaviour from the USP, UCH and OTU subfamilies, and use a combination of mutational analysis, mass spectrometry and structural studies to show that the inhibition is the result of the conversion of the enzymes' active-site cysteine to sulphenic acid (SOH) (Fig. 1).

For this modification to function as a reversible switch, further oxidation to sulphinic and sulphonic acid (SO_2H and SO_3H) must be avoided. The authors present two strategies by which this can be achieved. Lee *et al.* show that, for USP19, the oxidized cysteine bypasses further oxidation by reacting with a nitrogen atom in a neighbouring residue to generate a reversible sulphenylamide species (Fig. 1). Kulathu and colleagues' structural analysis of A20, an OTU enzyme with an SOH group at the active site, suggests that the architecture of the catalytic site provides the opportunity for hydrogen bonding that decreases further oxidative reactivity.

It is likely that many DUBs must continue to function under conditions of oxidative stress. How, then, do they avoid the negative influence of ROS? One suggestion comes from observations that some DUBs crystallize in inactive conformations. In these cases, the distance between the catalytic cysteine and histidine residues is too great for effective deprotonation. It is proposed that binding of the enzyme by its substrate or another molecule is required to realign the enzyme such that these residues come into their active configuration.

The regulation of ubiquitin-dependent signalling pathways by ROS was established[6] in a study which showed that treating cells with H_2O_2 enhances an inflammatory response involving the NF-κB signalling pathway. This coincided with ROS-dependent inhibition of the DUB enzyme Cezanne. Lee and colleagues sought further cellular manifestations of the influence of ROS on DUB-controlled processes. They showed that treating white blood cells called macrophages with H_2O_2, or stimulating them through Toll-like receptors to produce their own H_2O_2, leads to an overall reduction in cellular DUB activity. The authors also find that H_2O_2 treatment leads to the accumulation of the ubiquitinated form of PCNA, a protein that coordinates a DNA-repair pathway for mending damage caused by oxidative stress. Ubiquitination of PCNA, which is required for recruitment of repair enzymes, is held in check by the deubiquitinating activity of USP1. The accumulation of ubiquitinated PCNA suggests that H_2O_2 treatment results in direct inhibition of USP1. This finding mirrors those reported in another paper[7] that also generalizes the findings to other DUBs. Collectively, these papers[2,3,7] highlight the ubiquity of ROS sensitivity across the main cysteine-protease families of the DUBs.

Regulation of DUB activity by oxidation is probably important wherever the action of cysteine-protease DUBs coincides with excess ROS generation. Several DUBs have been implicated in growth-factor signalling pathways, and these must now be considered, along with phosphatases, as potential targets of ROS generation that follows stimulation with growth factors. However, a bone fide example of specific DUB regulation by intracellularly generated ROS awaits description. It is also worth considering that the addition of nitric oxide groups by reactive nitrogen species will similarly modify nucleophilic cysteine residues. Thus, it is possible that a similar means of control operates in physiological processes in which signalling by reactive nitrogen species and ubiquitin dynamics intersect, such as in the regulation of synapses between neurons. ■

Michael J. Clague *is in the Department of Cellular and Molecular Physiology, University of Liverpool, Liverpool L69 3BX, UK.*
e-mail: clague@liv.ac.uk

1. Paulsen, C. E. & Carroll, K. S. *ACS Chem. Biol.* **5,** 47–62 (2010).
2. Kulathu, Y. *et al. Nature Commun.* **4,** 1569 (2013).
3. Lee, J. G., Baek, K., Soetandyo, N. & Ye, Y. *Nature Commun.* **4,** 1568 (2013).
4. Tonks, N. K. *Cell* **121,** 667–670 (2005).
5. Komander, D., Clague, M. J. & Urbe, S. *Nature Rev. Mol. Cell Biol.* **10,** 550–563 (2009).
6. Enesa, K. *et al. J. Biol. Chem.* **283,** 18582–18590 (2008).
7. Cotto-Rios, X. M., Bekes, M., Chapman, J., Ueberheide, B. & Huang, T. T. *Cell Rep.* **2,** 1475–1484 (2012).

CORRECTION

In the News & Views article 'Solar System: Saturn's ring rain' by Jack Connerney (*Nature* **496,** 178–179; 2013), the unit of wavelength in Figure 2 was incorrectly given as millimetres. The correct unit is micrometres.

doi:10.1038/nature12047

Globally networked risks and how to respond

Dirk Helbing[1,2]

Today's strongly connected, global networks have produced highly interdependent systems that we do not understand and cannot control well. These systems are vulnerable to failure at all scales, posing serious threats to society, even when external shocks are absent. As the complexity and interaction strengths in our networked world increase, man-made systems can become unstable, creating uncontrollable situations even when decision-makers are well-skilled, have all data and technology at their disposal, and do their best. To make these systems manageable, a fundamental redesign is needed. A 'Global Systems Science' might create the required knowledge and paradigm shift in thinking.

Globalization and technological revolutions are changing our planet. Today we have a worldwide exchange of people, goods, money, information, and ideas, which has produced many new opportunities, services and benefits for humanity. At the same time, however, the underlying networks have created pathways along which dangerous and damaging events can spread rapidly and globally. This has increased systemic risks[1] (see Box 1). The related societal costs are huge.

When analysing today's environmental, health and financial systems or our supply chains and information and communication systems, one finds that these systems have become vulnerable on a planetary scale. They are challenged by the disruptive influences of global warming, disease outbreaks, food (distribution) shortages, financial crashes, heavy solar storms, organized (cyber-)crime, or cyberwar. Our world is already facing some of the consequences: global problems such as fiscal and economic crises, global migration, and an explosive mix of incompatible interests and cultures, coming along with social unrests, international and civil wars, and global terrorism.

In this Perspective, I argue that systemic failures and extreme events are consequences of the highly interconnected systems and networked risks humans have created. When networks are interdependent[2,3], this makes them even more vulnerable to abrupt failures[4–6]. Such interdependencies in our "hyper-connected world"[1] establish "hyper-risks" (see Fig. 1). For example, today's quick spreading of emergent epidemics is largely a result of global air traffic, and may have serious impacts on our global health, social and economic systems[6–9]. I also argue that initially beneficial trends such as globalization, increasing network densities, sparse use of resources, higher complexity, and an acceleration of institutional decision processes may ultimately push our anthropogenic (man-made or human-influenced) systems[10] towards systemic instability—a state in which things will inevitably get out of control sooner or later.

Many disasters in anthropogenic systems should not be seen as 'bad luck', but as the results of inappropriate interactions and institutional settings. Even worse, they are often the consequences of a wrong understanding due to the counter-intuitive nature of the underlying system behaviour. Hence, conventional thinking can cause fateful decisions and the repetition of previous mistakes. This calls for a paradigm shift in thinking: systemic instabilities can be understood by a change in perspective from a component-oriented to an interaction- and network-oriented view. This also implies a fundamental change in the design and management of complex dynamical systems.

The FuturICT community[11] (see http://www.futurict.eu), which involves thousands of scientists worldwide, is now engaged in establishing a 'Global Systems Science', in order to understand better our information society with its close co-evolution of information and communication technology (ICT) and society. This effort is allied with the "Earth system science"[10] that now provides the prevailing approach to studying the physics, chemistry and biology of our planet. Global Systems Science wants to make the theory of complex systems applicable to the solution of global-scale problems. It will take a massively data-driven approach that builds on a serious collaboration between the natural, engineering, and social sciences, aiming at a grand integration of knowledge. This approach to real-life techno-socio-economic-environmental systems[8] is expected to enable new response strategies to a number of twenty-first century challenges.

BOX 1

Risk, systemic risk and hyper-risk

According to the standard ISO 31000 (2009; http://www.iso.org/iso/catalogue_detail?csnumber=43170), risk is defined as "effect of uncertainty on objectives". It is often quantified as the probability of occurrence of an (adverse) event, times its (negative) impact (damage), but it should be kept in mind that risks might also create positive impacts, such as opportunities for some stakeholders.

Compared to this, systemic risk is the risk of having not just statistically independent failures, but interdependent, so-called 'cascading' failures in a network of N interconnected system components. That is, systemic risks result from connections between risks ('networked risks'). In such cases, a localized initial failure ('perturbation') could have disastrous effects and cause, in principle, unbounded damage as N goes to infinity. For example, a large-scale power blackout can hit millions of people. In economics, a systemic risk could mean the possible collapse of a market or of the whole financial system. The potential damage here is largely determined by the size N of the networked system.

Even higher risks are implied by networks of networks[4,5], that is, by the coupling of different kinds of systems. In fact, new vulnerabilities result from the increasing interdependencies between our energy, food and water systems, global supply chains, communication and financial systems, ecosystems and climate[10]. The World Economic Forum has described this situation as a hyper-connected world[1], and we therefore refer to the associated risks as 'hyper-risks'.

[1]ETH Zurich, Clausiusstrasse 50, 8092 Zurich, Switzerland. [2]Risk Center, ETH Zurich, Swiss Federal Institute of Technology, Scheuchzerstrasse 7, 8092 Zurich, Switzerland.

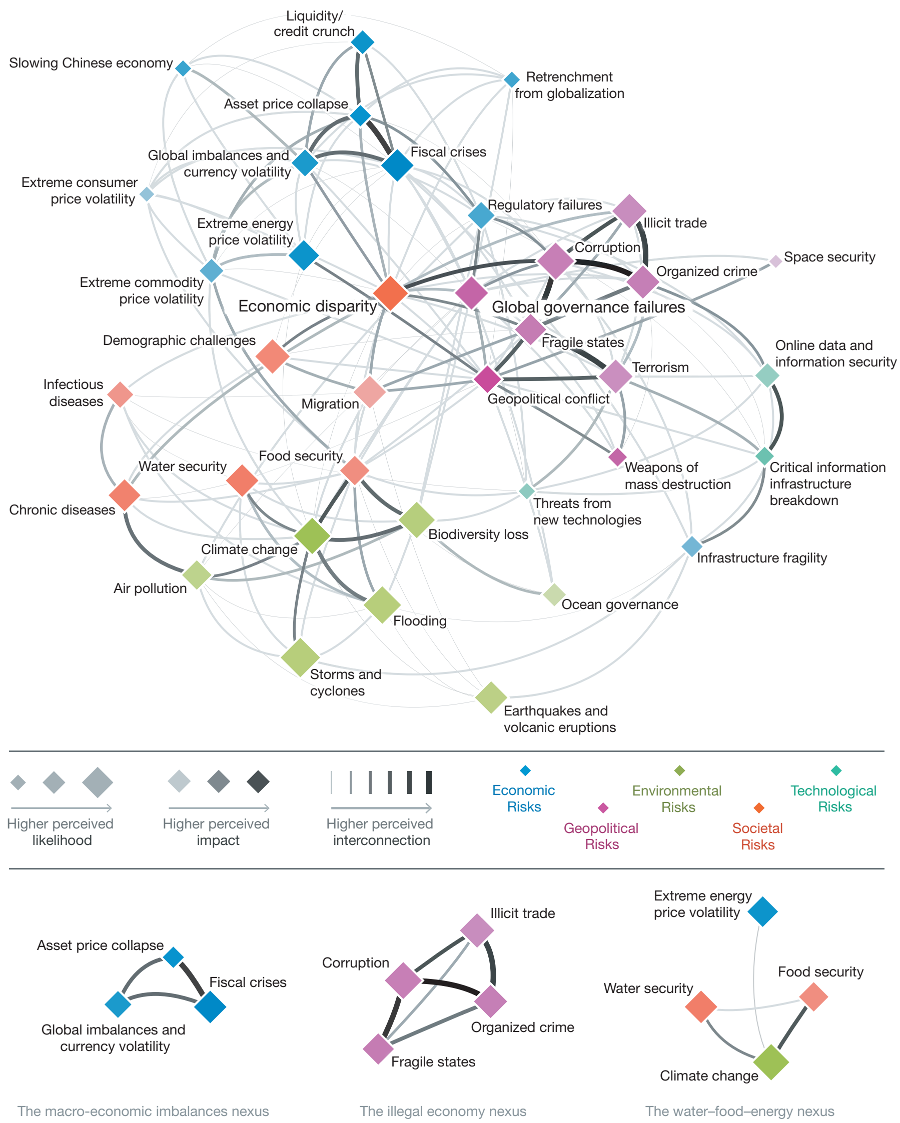


Figure 1 | Risks Interconnection Map 2011 illustrating systemic interdependencies in the hyper-connected world we are living in. Reprinted from ref. 82 with permission of the WEF.

What we know

Overview

Catastrophe theory[12] suggests that disasters may result from discontinuous transitions in response to gradual changes in parameters. Such systemic shifts are expected to occur at certain 'tipping points' (that is, critical parameter values) and lead to different system properties. The theory of critical phenomena[13] has shown that, at such tipping points, power-law (or other heavily skewed) distributions of event sizes are typical. They relate to cascade effects[4,5,14–20], which may have any size. Hence, "extreme events"[21] can be a result of the inherent system dynamics rather than of unexpected external events. The theory of self-organized criticality[22] furthermore shows that certain systems (such as piles of grains prone to avalanches) may be automatically driven towards a critical tipping point. Other work has studied the error and attack tolerance of networks[23] and cascade effects in networks[4,5,14–20,24], where local failures of nodes or links may trigger overloads and consequential failures of other nodes or links. Moreover, abrupt systemic failures may result from interdependencies between networks[4–6] or other mechanisms[25,26].

Surprising behaviour due to complexity

Current anthropogenic systems show an increase of structural, dynamic, functional and algorithmic complexity. This poses challenges for their design, operation, reliability and efficiency. Here I will focus on complex dynamical systems—those that cannot be understood by the sum of their components' properties, in contrast to loosely coupled systems. The following typical features result from the nonlinear interactions in complex systems[27,28]. (1) Rather than having one equilibrium solution, the system might show numerous different behaviours, depending on the respective initial conditions. (2) Complex dynamical systems may seem uncontrollable. In particular, opportunities for external or top-down control are very limited[29]. (3) Self-organization and strong correlations dominate the system behaviour. (4) The (emergent) properties of complex dynamical systems are often surprising and counter-intuitive[30].

Furthermore, the combination of nonlinear interactions, network effects, delayed response and randomness may cause a sensitivity to small changes, unique path dependencies, and strong correlations, all of which are hard to understand, prepare for and manage. Each of these factors is already difficult to imagine, but this applies even more to their combination.

For example, fundamental changes in the system outcome—such as non-cooperative behaviour rather than cooperation among agents—can result from seemingly small changes in the nature of the components or their mode of interaction (see Fig. 2). Such small changes may be interactions that take place on particular networks rather than on regular or random networks, interactions or components that are spatially varying rather than homogeneous, or which are subject to random 'noise' rather than behaving deterministically[31,32].

Cascade effects due to strong interactions

Our society is entering a new era—the era of a global information society, characterized by increasing interdependency, interconnectivity and complexity, and a life in which the real and digital world can no longer be separated (see Box 2). However, as interactions between components become 'strong', the behaviour of system components may seriously alter or impair the functionality or operation of other components. Typical properties of strongly coupled systems in the above-defined sense are: (1) Dynamical changes tend to be fast, potentially outstripping the rate at which one can learn about the characteristic system behaviour, or at which humans can react. (2) One event can trigger further events, thereby creating amplification and cascade effects[4,5,14–20], which implies a large vulnerability to perturbations, variations or random failures. Cascade effects come along with highly correlated transitions of many system components or variables from a stable to an unstable state, thereby driving the system out of equilibrium. (3) Extreme events tend to occur more often than expected for normally distributed event sizes[17,21].

Probabilistic cascade effects in real-life systems are often hard to identify, understand and map. Rather than deterministic one-to-one relationships between 'causes' and 'effects', there are many possible paths of events (see Fig. 3), and effects may occur with obfuscating delays.

Systemic instabilities challenge our intuition

Why are attempts to control strongly coupled, complex systems so often unsuccessful? Systemic failures may occur even if everybody involved is highly skilled, highly motivated and behaving properly. I shall illustrate this with two examples.

Crowd disasters

Crowd disasters constitute an eye-opening example of the eventual failure of control in a complex system. Even if nobody wants to harm anybody else, people may be fatally injured. A detailed analysis reveals amplifying feedback effects that cause a systemic instability[33,34]. The interaction strength increases with the crowd density, as people come closer together. When the density becomes too high, inadvertent contact forces are transferred from one body to another and add up. The resulting forces vary significantly in direction and size, pushing people around, and creating a phenomenon called 'crowd quake'. Turbulent waves cause people to stumble, and others fall over them in an often fatal domino effect. If people do not manage to get back on their feet quickly enough, they are likely to suffocate. In many cases, the instability is created not by foolish or malicious individual actions, but by the unavoidable amplification of small fluctuations above a critical density threshold. Consequently, crowd disasters cannot simply be evaded by policing, aimed at imposing 'better behaviour'. Some kinds of crowd control might even worsen the situation[34].

Financial meltdown

Almost a decade ago, the investor Warren Buffett warned that massive trade in financial derivatives would create mega-catastrophic risks for the economy. In the same context, he spoke of an investment "time bomb" and of financial derivatives as "weapons of mass destruction" (see http://news.bbc.co.uk/2/hi/2817995.stm, accessed 1 June 2012). Five years later, the financial bubble imploded and destroyed trillions of stock value. During this time, the overall volume of credit default swaps and other financial derivatives had grown to several times the world gross domestic product.

But what exactly caused the collapse? In response to the question by the Queen of England of why nobody had foreseen the financial crisis, the British Academy concluded: "Everyone seemed to be doing their own job properly on its own merit. And according to standard measures of success, they were often doing it well. The failure was to see how collectively this added up to a series of interconnected imbalances... Individual risks may rightly have been viewed as small, but the risk to the system as a whole was vast." (See http://www.britac.ac.uk/templates/asset-relay.cfm?frmAssetFileID=8285, accessed 1 June 2012.) For example,

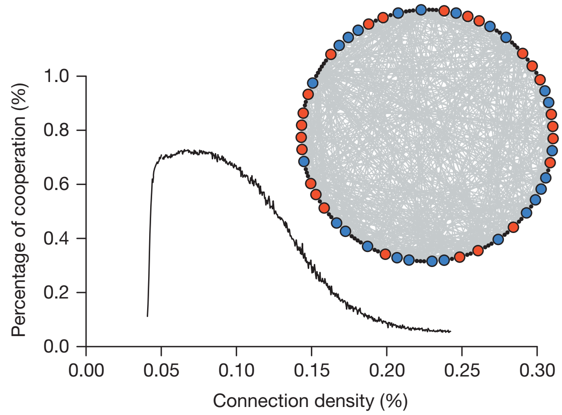


Figure 2 | Spreading and erosion of cooperation in a prisoner's dilemma game. The computer simulations assume the payoff parameters $T = 7$, $R = 6$, $P = 2$, and $S = 1$ and include success-driven migration[32]. Although cooperation would be profitable to everyone, non-cooperators can achieve a higher payoff than cooperators, which may destabilize cooperation. The graph shows the fraction of cooperative agents, averaged over 100 simulations, as a function of the connection density (actual number of network links divided by the maximum number of links when all nodes are connected to all others). Initially, an increasing link density enhances cooperation, but as it passes a certain threshold, cooperation erodes. (See http://vimeo.com/53876434 for a related movie.) The computer simulations are based on a circular network with 100 nodes, each connected with the four nearest neighbours. n links are added randomly. 50 nodes are occupied by agents. The inset shows a 'snapshot' of the system: blue circles represent cooperation, red circles non-cooperative behaviour, and black dots empty sites. Initially, all agents are non-cooperative. Their network locations and behaviours (cooperation or defection) are updated in a random sequential way in 4 steps: (1) The agent plays two-person prisoner's dilemma games with its direct neighbours in the network. (2) After the interaction, the agent moves with probability 0.5 up to 4 steps along existing links to the empty node that gives the highest payoff in a fictitious play step, assuming that noone changes the behaviour. (3) The agent imitates the behaviour of the neighbour who got the highest payoff in step 1 (if higher than the agent's own payoff). (4) The behaviour is spontaneously changed with a mutation rate of 0.1.

while risk diversification in a banking system is aimed at minimizing risks, it can create systemic risks when the network density becomes too high[20].

Drivers of systemic instabilities

Table 1 lists common drivers of systemic instabilities[32], and what makes the corresponding system behaviours difficult to understand. Current global trends promote several of these drivers. Although they often have desirable effects in the beginning, they may destabilize anthropogenic systems over time. Such drivers are, for example: (1) increasing system sizes, (2) reduced redundancies due to attempts to save resources (implying a loss of safety margins), (3) denser networks (creating increasing interdependencies between critical parts of the network, see Figs 2 and 4), and (4) a high pace of innovation[35] (producing uncertainties or 'unknown unknowns'). Could these developments create a "global time bomb"? (See Box 3.)

Knowledge gaps

Not well behaved

The combination of complex interactions with strong couplings can lead to surprising, potentially dangerous system behaviours[17,30], which are barely understood. At present, most of the scientific understanding of large networks is restricted to cases of special, sparse, or static networks. However, dynamically changing, strongly coupled, highly interconnected and densely populated complex systems are fundamentally different[36]. The number of possible system behaviours and proper management strategies, when regular interaction networks are replaced by irregular ones, is overwhelming[18]. In other words, there is no standard solution for complex systems, and 'the devil is in the detail'.

BOX 2

Global information and communication systems

One vulnerable system deserving particular attention is our global network of information and communication technologies (ICT)[11]. Although these technologies will be central to the solution of global challenges, they are also part of the problem and raise fundamental ethical issues, for example, how to ensure the self-determined use of personal data. New 'cyber-risks' arise from the fact that we are now enormously dependent on reliable information and communication systems. This includes threats to individuals (such as privacy intrusion, identity theft or manipulation by personalized information), to companies (such as cybercrime), and to societies (such as cyberwar or totalitarian control).

Our global ICT system is now the biggest artefact ever created, encompassing billions of diverse components (computers, smartphones, factories, vehicles and so on). The digital and real world cannot be divided any more; they form a single interweaved system. In this new "cybersocial world", digital information drives real events. The techno-socio-economic implications of all this are barely understood[11]. The extreme speed of these systems, their hyper-connectivity, large complexity, and massive data volumes produced are often seen as problems. Moreover, the components increasingly make autonomous decisions. For example, supercomputers are now performing the majority of financial transactions. The 'flash crash' of 6 May 2010 illustrates the unexpected systemic behaviour that can result (http://en.wikipedia.org/wiki/2010_Flash_Crash, accessed 29 July 2012): within minutes, nearly $1 trillion in market value disappeared before the financial markets recovered again. Such computer systems can be considered to be 'artificial social systems', as they learn from information about their environment, develop expectations about the future, and decide, interact and communicate autonomously. To design these systems properly, ensure a suitable response to human needs, and avoid problems such as co-ordination failures, breakdowns of cooperation, conflict, (cyber-)crime or (cyber-)war, we need a better, fundamental understanding of socially interactive systems.

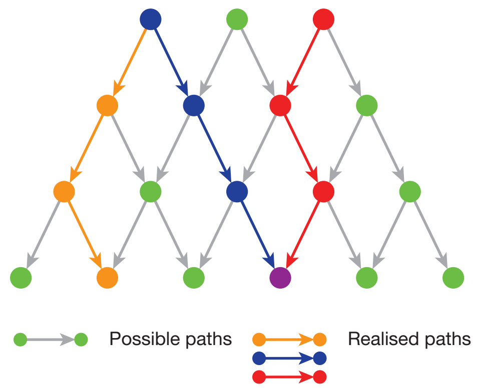


Figure 3 | Illustration of probabilistic cascade effects in systems with networked risks. The orange and blue paths show that the same cause can have different effects, depending on the respective random realization. The blue and red paths show that different causes can have the same effect. The understanding of cascade effects requires knowledge of at least the following three contributing factors: the interactions in the system, the context (such as institutional or boundary conditions), and in many cases, but not necessarily so, a triggering event (i.e. randomness may determine the temporal evolution of the system). While the exact timing of the triggering event is often not predictable, the post-trigger dynamics might be foreseeable to a certain extent (in a probabilistic sense). When system components behave randomly, a cascade effect might start anywhere, but the likelihood to originate at a weak part of the system is higher (e.g. traffic jams mostly start at known bottlenecks, but not always).

Moreover, most existing theories do not provide much practical advice on how to respond to actual global risks, crises and disasters, and empirically based risk-mitigation strategies often remain qualitative[37–42]. Most scientific studies make idealized assumptions such as homogeneous components, linear, weak or deterministic interactions, optimal and independent behaviours, or other favourable features that make systems well-behaved (smooth dependencies, convex sets, and so on). Real-life systems, in contrast, are characterized by heterogeneous components, irregular interaction networks, nonlinear interactions, probabilistic behaviours, interdependent decisions, and networks of networks. These differences can change the resulting system behaviour fundamentally and dramatically and in unpredictable ways. That is, real-world systems are often not well-behaved.

Behavioural rules may change

Many existing risk models also neglect the special features of social systems, for example, the importance of a feedback of the emergent macro-level dynamics on the micro-level behaviour of the system components or on specific information input (see Box 4). Now, a single video or tweet may cause deadly social unrest on the other side of the globe. Such changes of the microdynamics may also change the failure probabilities of system components.

For example, consider a case in which interdependent system components may fail or not with certain probabilities, and where local damage increases the likelihood of further damage. As a consequence, the bigger a failure cascade, the higher the probability that it might grow larger. This establishes the possibility of global catastrophic risks (see Fig. 4), which cannot be reasonably insured against. The decreasing capacity of a socio-economic system to recover as a cascade failure progresses (thereby eliminating valuable resources needed for recovery) calls for a strong effort to stop cascades right at the beginning, when the damage is still small and the problem may not even be perceived as threatening. Ignoring this important point may cause costly and avoidable damage.

Fundamental and man-made uncertainty

Systems involving uncertainty, where the probability of particular events (for example, the occurrence of damage of a certain size) cannot be specified, are probably the least understood. Uncertainty may be a result of limitations of calibration procedures or lack of data. However, it may also have a fundamental origin. Let us assume a system of systems, in which the output variables of one system are input variables of another one. Let us further assume that the first system is composed of well-behaved components, whose variables are normally distributed around their equilibrium state. Connecting them strongly may nevertheless cause cascade effects and power-law-distributed output variables[13]. If the exponent of the related cumulative distribution function is between -2 and -1, the standard deviation is not defined, and if it is between -1 and 0, not even the mean value exists. Hence, the input variables of the second system could have any value, and the damage in the second system depends on the actual, unpredictable values of the input variables. Then, even if one had all the data in the world, it would be impossible to predict or control the outcome. Under such conditions it is not possible to protect the system from catastrophic failure. Such problems must and can only be solved by a proper (re)design of the system and suitable management principles, as discussed in the following.

Some design and operation principles

Managing complexity using self-organization

When systems reach a certain size or level of complexity, algorithmic constraints often prohibit efficient top-down management by real-time optimization. However, "guided self-organisation"[32,43,44] is a promising alternative way of managing complex dynamical systems, in a decentralized, bottom-up way. The underlying idea is to use, rather than fight, the system-immanent tendency of complex systems to self-organize and thereby create a stable, ordered state. For this, it is important to have the

Table 1 | Drivers and examples of systemic instabilities

Driver/factor	Description/phenomenon	Field/modelling approach	Examples	Surprising system behaviour
Threshold effect	Unexpected transition, systemic shift	Bifurcation[73] and catastrophe theory[12], explosive percolation[25], dragon kings[26]	Revolutions (for example, the Arab Spring, breakdown of former GDR, now East Germany)	Sudden failure of continuous improvement attempts
Randomness in a strongly coupled system	Strong correlations, mean-field approximation ('representative agent model') does not work	Statistical physics, theory of critical phenomena[13]	Self-organized criticality[22], earthquakes[74], stock market variations, evolutionary jumps, floods, sunspots	Extreme events[21], outcome can be opposite of mean-field prediction
Positive feedback	Dynamic instability and amplification effect, equilibrium or stationary state cannot be maintained	(Linear) stability analysis, eigenvalues theory, sensitivity analysis	Tragedy of the commons[31] (tax evasion, over-fishing, exploitation of environment, global warming, free-riding, misuse of social benefits)	Bubbles and crashes, cooperation breaks down, although it would be better for everyone
Wrong timing (mismatch of adjustment processes)	Over-reaction, growing oscillations, loss of synchronization[51]	(Linear) stability analysis, eigenvalue theory	Phantom traffic jams[75], blackout of electrical power grids[76]	Breakdown of flow despite sufficient capacity
Strong interaction, contagion	Domino and cascade effects, avalanches	Network analysis, agent-based models, bundle-fibre model[24]	Financial crisis, epidemic spreading[8]	It may be impossible to enumerate the risk
Complex structure	Perturbations in one network affect another one	Theory of interdependent networks[4]	Coupled electricity and communication networks, impact of natural disasters on critical infrastructures	Possibility of sudden failure (rather than gradual deterioration of performance)
Complex dynamics	Self-organized dynamics, emergence of new systemic properties	Nonlinear dynamics, chaos theory[77], complexity theory[28]	Crowd turbulence[33]	Systemic properties differ from the component properties
Complex function	Sensitivity, opaqueness, scientific unknowns	Computational and experimental testing	Information and communication systems	Unexpected system properties and failures
Complex control	Time required for computational solution explodes with system size, delayed or non-optimal solutions	Cybernetics[78], heuristics	Traffic light control[45], production, politics	Optimal solution unreachable, slower-is-faster effect[75]
Optimization	Orientation at state of high performance; loss of reserves and redundancies	Operations research	Throughput optimization, portfolio optimization	Capacity drop[75], systemic risks created by insurance against risks[79]
Competition	Incompatible preferences or goals	Economics, political sciences	Conflict[72]	Market failure, minority may win
Innovation	Introduction of new system components, designs or properties; structural instability[80]	Evolutionary models, genetic algorithms[68]	Financial derivatives, new products, new procedures and new species	Point change can mess up the whole system, finite time singularity[35,81]

right kinds of interactions, adaptive feedback mechanisms, and institutional settings. By establishing proper 'rules of the game', within which the system components can self-organize, including mechanisms ensuring rule compliance, top-down and bottom-up principles can be combined and inefficient micro-management can be avoided. To overcome suboptimal solutions and systemic instabilities, the interaction rules or institutional settings may have to be modified. Symmetrical interactions, for example, can often promote a well-balanced situation and an evolution to the optimal system state[32].

Traffic light control is a good example to illustrate the ongoing paradigm shift in managing complexity. Classical control is based on the principle of a 'benevolent dictator': a traffic control centre collects information from the city and tries to impose an optimal traffic light control. But because the optimization problem is too demanding for real-time optimization, the control scheme is adjusted for the typical traffic flows on a certain day and time. However, this control is not optimal for the actual situation owing to the large variability in the arrival rates of vehicles.

Significantly smaller and more predictable travel times can be reached using a flexible "self-control" of traffic flows[45]. This is based on a suitable real-time response to a short-term anticipation of vehicle flows, thereby coordinating neighbouring intersections. Decentralized principles of managing complexity are also used in information and communication systems[46], and they are becoming a trend in energy production ("smart grids"[47]). Similar self-control principles could be applied to logistic and production systems, or even to administrative processes and governance.

Coping with networked risks

To cope with hyper-risks, it is necessary to develop risk competence and to prepare and exercise contingency plans for all sorts of possible failure cascades[4,5,14–20]. The aim is to attain a resilient ('forgiving') system design and operation[48,49].

An important principle to remember is to have at least one backup system that runs in parallel to the primary system and ensures a safe fallback level. Note that a backup system should be operated and designed according to different principles in order to avoid a failure of both systems for the same reasons. Diversity may not only increase systemic resilience (that is, the ability to absorb shocks or recover from them), it can also promote systemic adaptability and innovation[43]. Furthermore, diversity makes it less likely that all system components fail at the same time. Consequently, early failures of weak system components (critical fluctuations) will create early warning signals of an impending systemic instability[50].

An additional principle of reducing hyper-risks is the limitation of system size, to establish upper bounds to the possible scale of disaster. Such a limitation might also be established in a dynamical way, if real-time feedback allows one to isolate affected parts of the system before others are damaged by cascade effects. If a sufficiently rapid dynamic decoupling cannot be ensured, one can build weak components (breaking points) into the system, preferably in places where damage would be comparatively small. For example, fuses in electrical circuits serve to avoid large-scale damage of local overloads. Similarly, engineers have learned to build crush zones in cars to protect humans during accidents.

A further principle would be to incorporate mechanisms producing a manageable state. For example, if the system dynamics unfolds so rapidly that there is a danger of losing control, one could slow it down by introducing frictional effects (such as a financial transaction fee that kicks in when financial markets drop).

Also note that dynamical processes in a system can desynchronize[51], if the control variables change too quickly relative to the timescale on which the governed components can adjust. For example, stable hierarchical systems typically change slowly on the top and much quicker on the lower levels. If the influence of the top on the bottom levels becomes

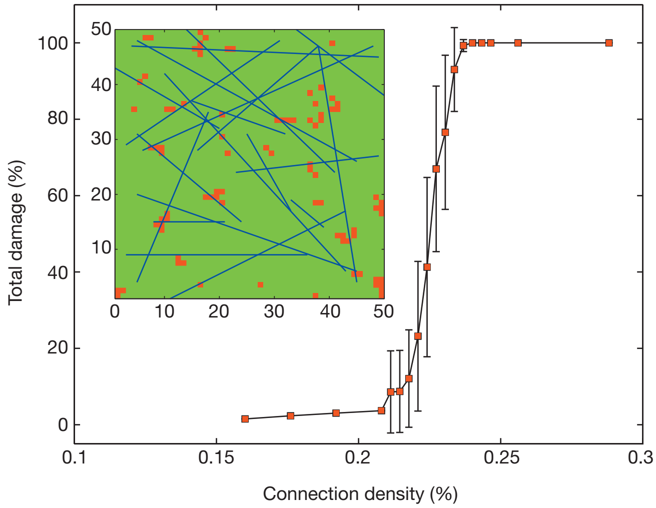


Figure 4 | Cascade spreading is increasingly hard to recover from as failure progresses. The simulation model mimics spatial epidemic spreading with air traffic and healing costs in a two-dimensional 50×50 grid with periodic boundary conditions and random shortcut links. The colourful inset depicts an early snapshot of the simulation with $N = 2{,}500$ nodes. Red nodes are infected, green nodes are healthy. Shortcut links are shown in blue. The connectivity-dependent graph shows the mean value and standard deviation of the fraction $i(t)/N$ of infected nodes over 50 simulation runs. Most nodes have four direct neighbours, but a few of them possess an additional directed random connection to a distant node. The spontaneous infection rate is $s = 0.001$ per time step; the infection rate by an infected neighbouring node is $P = 0.08$. Newly infected nodes may infect others or may recover from the next time step onwards. Recovery occurs with a rate $q = 0.4$, if there is enough budget $b > c$ to bear the healing costs $c = 80$. The budget needed for recovery is created by the number of healthy nodes $h(t)$. Hence, if $r(t)$ nodes are recovering at time t, the budget changes according to $b(t + 1) = b(t) + h(t) - cr(t)$. As soon as the budget is used up, the infection spreads explosively. (See also the movie at http://vimeo.com/53872893.)

too strong, this may impair the functionality and self-organization of the hierarchical structure[32].

Last but not least, reducing connectivity may serve to decrease the coupling strength in the system. This implies a change from a dense to a sparser network, which can reduce contagious spreading effects. In fact, sparse networks seem to be characteristic for ecological systems[52].

As logical as the above safety principles may sound, these precautions have often been neglected in the design and operation of strongly coupled, complex systems such as the world financial system[20,53,54].

What is ahead

Despite all our knowledge, much work is still ahead of us. For example, the current financial crisis shows that much of our theoretical knowledge has not yet found its way into real-world policies, as it should.

Economic crises

Two main pillars of mainstream economics are the equilibrium paradigm and the representative agent approach. According to the equilibrium paradigm, economies are viewed as systems that tend to evolve towards an equilibrium state. Bubbles and crashes should not happen and, hence, would not require any precautions[54]. Sudden changes would be caused exclusively by external shocks. However, it does not seem to be widely recognized that interactions between system elements can cause amplifying cascade effects even if all components relax to their equilibrium state[55,56].

Representative agent models, which assume that companies act in the way a representative (average) individual would optimally decide, are more general and allow one to describe dynamical processes. However, such models cannot capture processes well if random events, the diversity of system components, the history of the system or correlations between variables matter a lot. It can even happen that representative

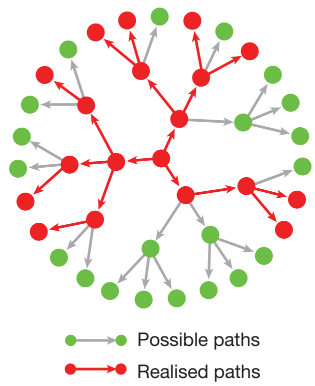
BOX 3

Have humans created a 'global time bomb'?

For a long time, crowd disasters and financial crashes seemed to be puzzling, unrelated, 'God-given' phenomena one simply had to live with. However, it is possible to grasp the mechanisms that cause complex systems to get out of control. Amplification effects can result and promote failure cascades, when the interactions of system components become stronger than the frictional effects or when the damaging impact of impaired system components on other components occurs faster than the recovery to their normal state.

For certain kinds of interaction networks, the similarity of related cascade effects with those of chain reactions in nuclear fission is disturbing (see Box 3 Figure). It is known that such processes are difficult to control. Catastrophic damage is a realistic scenario. Given the similarity of the cascading mechanisms, is it possible that our worldwide anthropogenic system will get out of control sooner or later? In other words, have humans unintentionally created something like a "global time bomb"?

If so, what kinds of global catastrophic scenarios might humans in complex societies[81] face? A collapse of the global information and communication systems or of the world economy? Global pandemics[6–9]? Unsustainable growth, demographic or environmental change? A global food or energy crisis? The large-scale spreading of toxic substances? A cultural clash[83]? Another global-scale conflict[84,85]? Or, more likely, a combination of several of these contagious phenomena (the "perfect storm"[1])? When analysing such global risks, one should bear in mind that the speed of destructive cascade effects might be slow, and the process may not look like an explosion. Nevertheless, the process can be hard to stop. For example, the dynamics underlying crowd disasters is slow, but deadly.



Box 3 Figure | Illustration of the principle of a 'time bomb'. A single, local perturbation of a node may cause large-scale damage through a cascade effect, similar to chain reactions in nuclear fission.

agent models make predictions opposite to those of agent-based computer simulations assuming the very same interaction rules[32] (see Fig. 2).

Paradigm shift ahead

Both equilibrium and representative agent models are fundamentally incompatible with probabilistic cascade effects—they are different classes of models. Cascade effects cause a system to leave its previous (equilibrium) state, and there is also no representative dynamics, because different possible paths of events may look very different (see Fig. 3). Considering furthermore that the spread of innovations and products also involves cascade effects[57,58], it seems that cascade effects are even the rule rather than the exception in today's economy. This calls for a new economic thinking. Many currently applied theories are based on the

BOX 4

Social factors and social capital

Many twenty-first-century challenges have a social component and cannot be solved by technology alone[86]. Socially interactive systems, be it social or economic systems, artificial societies, or the hybrid system made up of our virtual and real worlds, are characterized by a number of special features, which imply additional risks: The components (for example, individuals) take autonomous decisions based on (uncertain) future expectations. They produce and respond to complex and often ambiguous information. They have cognitive complexity. They have individual learning histories and therefore different, subjective views of reality. Individual preferences and intentions are diverse, and imply conflicts of interest. The behaviour may depend on the context in a sensitive way. For example, the way people behave and interact may change in response to the emergent social dynamics on the macro scale. This also implies the ability to innovate, which may create surprising outcomes and 'unknown unknowns' through new kinds of interactions. Furthermore, social network interactions can create social capital[43,87] such as trust, solidarity, reliability, happiness, social values, norms and culture.

To assess systemic risks fully, a better understanding of social capital is crucial. Social capital is important for economic value generation, social well-being, and societal resilience, but it may be damaged or exploited, like our environment. Therefore, humans need to learn how to quantify and protect social capital[36]. A warning example is the loss of trillions of dollars in the stock markets during the financial crisis, which was largely caused by a loss of trust. It is important to stress that risk insurances today do not consider damage to social capital. However, it is known that large-scale disasters have a disproportionate public impact, which is related to the fact that they destroy social capital. By neglecting social capital in risk assessment, we are taking higher risks than we would rationally do.

assumption that statistically independent, optimal decisions are made. Under such idealized conditions one can show that financial markets are efficient, that herding effects will not occur, and that unregulated, self-regarding behaviour can maximize system performance, benefiting everyone. Some of these paradigms are centuries old yet still applied by policy-makers. However, such concepts must be questioned in a world where economic decisions are strongly coupled and cascade effects are frequent[54,59].

Global Systems Science

For a long time, humans have considered systemic failures to originate from 'outside the system', because it has been difficult to understand how they could come about otherwise. However, many disasters in anthropogenic systems result from a wrong way of thinking and, consequently, from inappropriate organization and systems design. For example, we often apply theories for well-behaved systems to systems that are not well behaved.

Given that many twenty-first-century problems involve socio-economic challenges, we need to develop a science of economic systems that is consistent with our knowledge of complex systems. A massive interdisciplinary research effort is indispensable to accelerate science and innovation so that our understanding and capabilities can keep up with the pace at which our world is changing ('innovation acceleration'[11]).

In the following, I use the term Global Systems Science to emphasize that integrating knowledge from the natural, engineering and social sciences and applying it to real-life systems is a major challenge that goes beyond any currently existing discipline. There are still many unsolved problems regarding the interplay between structure, dynamics and functional properties of complex systems. A good overview of global interdependencies between different kinds of networks is lacking as well. The establishment of a Global Systems Science should fill these knowledge gaps, particularly regarding the role of human and social factors.

BOX 5

Beyond current risk analysis

State-of-the-art risk analysis[88] still seems to have a number of shortcomings. (1) Estimates for the probability distribution and parameters describing rare events, including the variability of such parameters over time, are often poor. (2) The likelihood of coincidences of multiple unfortunate, rare events is often underestimated (but there is a huge number of possible coincidences). (3) Classical fault tree and event tree analyses[37] (see also http://en.wikipedia.org/wiki/Fault tree analysis and http://en.wikipedia.org/wiki/Event tree, both accessed 18 November 2012) do not sufficiently consider feedback loops. (4) The combination of probabilistic failure analysis with complex dynamics is still uncommon, even though it is important to understand amplification effects and systemic instabilities. (5) The relevance of human factors, such as negligence, irresponsible or irrational behaviour, greed, fear, revenge, perception bias, or human error is often underestimated[30,41]. (6) Social factors, including the value of social capital, are typically not considered. (7) Common assumptions underlying established ways of thinking are not questioned enough, and attempts to identify uncertainties or 'unknown unknowns' are often insufficient. Some of the worst disasters have happened because of a failure to imagine that they were possible[42], and thus to guard against them. (8) Economic, political and personal incentives are not sufficiently analysed as drivers of risks. Many risks can be revealed by looking for stakeholders who could potentially profit from risk-taking, negligence or crises. Risk-seeking strategies that attempt to create new opportunities via systemic change are expected mainly under conditions of uncertainty, because these tend to be characterized by controversial debates and, therefore, under-regulation.

To reach better risk assessment and risk reduction we need transparency, accountability, responsibility and awareness of individual and institutional decision-makers[11,36]. Modern governance sometimes dilutes responsibility so much that nobody can be held responsible anymore and catastrophic risks may be a consequence. The financial crisis seems to be a good example. Part of the problem appears to be that credit default swaps and other financial derivatives are modern financial insurance instruments, which transfer risks from the individuals or institutions causing them to others, thereby encouraging excessive risk taking. It might therefore be necessary to establish a principle of collective responsibility, by which individuals or institutions share responsibility for incurred damage in proportion to their previous (and subsequent) gains.

Progress must be made in computational social science[60], for example by performing agent-based computer simulations[32,61–63] of learning agents with cognitive abilities and evolving properties. We also require the close integration of theoretical and computational with empirical and experimental efforts, including interactive multi-player serious games[64,65], laboratory and web experiments, and the mining of large-scale activity data[11].

We furthermore lack good methods of calculating networked risks. Modern financial derivatives package many risks together. If the correlations between the components' risks are stable in time, copula methodology[66] offers a reasonable modelling framework. However, the correlations strongly depend on the state of the global financial system[67]. Therefore, we still need to learn how realistically to calculate the interdependence and propagation of risks in a network, how to absorb them, and how to calibrate the models (see Box 5). This requires the integration of probability calculus, network theory and complexity science with large-scale data mining.

Making progress towards a better understanding of complex systems and systemic risks also depends crucially on the collection of 'big data' (massive amounts of data) and the development of powerful machine learning techniques that allow one to develop and validate realistic

explanatory models of interdependent systems. The increasing availability of detailed activity data and of cheap, ubiquitous sensing technologies will enable previously unimaginable breakthroughs.

Finally, given that it can be dangerous to introduce new kinds of components, interactions or interdependencies into our global systems, a science of integrative systems design is needed. It will have to elaborate suitable interaction rules and system architectures that ensure not only system components to work well, but also favourable systemic interactions and outcomes. A particular challenge is to design value-sensitive information systems and financial exchange systems that promote awareness and responsible action[11]. How could we create open information platforms that minimize misuse? How could we avoid privacy intrusion and the manipulation of individuals? How could we enable greater participation of citizens in social, economic and political affairs?

Finding tailored design and operation principles for complex, strongly coupled systems is challenging. However, inspiration can be drawn from ecological[52], immunological[68], and social systems[32]. Understanding the principles that make socially interactive systems work well (or not) will facilitate the invention of a whole range of socio-inspired design and operation principles[11]. This includes reputation, trust, social norms, culture, social capital and collective intelligence, all of which could help to counter cybercrime and to design a trustable future Internet.

New exploration instruments

To promote Global Systems Science with its strong focus on interactions and global interdependencies, the FuturICT initiative proposes to build new, open exploration instruments ('socioscopes'), analogous to the telescopes developed earlier to explore new continents and the universe. One such instrument, called the "Planetary Nervous System"[11], would process data reflecting the state and dynamics of our global techno-socio-economic-environmental system. Internet data combined with data collected by sensor networks could be used to measure the state of our world in real time[69]. Such measurements should reflect not only physical and environmental conditions, but also quantify the "social footprint"[11], that is, the impact of human decisions and actions on our socio-economic system. For example, it would be desirable to develop better indices of social wellbeing than the gross domestic product per capita, ones that consider environmental factors, health and human and social capital (see Box 4 and http://www.stiglitz-sen-fitoussi.fr and http://www.worldchanging.com/archives/010627.html). The Planetary Nervous System would also increase collective awareness of possible problems and opportunities, and thereby help us to avoid mistakes.

The data generated by the Planetary Nervous System could be used to feed a "Living Earth Simulator"[11], which would simulate simplified, but sufficiently realistic models of relevant aspects of our world. Similar to weather forecasts, an increasingly accurate picture of our world and its possible evolutions would be obtained over time as we learn to model anthropogenic systems and human responses to information. Such 'policy wind tunnels' would help to analyse what-if scenarios, and to identify strategic options and their possible implications. This would provide a new tool with which political decision-makers, business leaders, and citizens could gain a better, multi-perspective picture of difficult matters.

Finally, a "Global Participatory Platform"[11] would make these new instruments accessible to everybody and create an open 'information ecosystem', which would include an interactive platform for crowd sourcing and cooperative applications. The activity data generated there would also allow one to determine statistical laws of human decision making and collective action[64]. Furthermore, it would be conceivable to create interactive virtual worlds[65] in order to explore possible futures (such as alternative designs of urban areas, financial architectures and decision procedures).

Discussion

I have described how system components, even if their behaviour is harmless and predictable when separated, can create unpredictable and uncontrollable systemic risks when tightly coupled together. Hence, an improper design or management of our global anthropogenic system creates possibilities of catastrophic failures.

Today, many necessary safety precautions to protect ourselves from human-made disasters are not taken owing to insufficient theoretical understanding and, consequently, wrong policy decisions. It is dangerous to believe that crises and disasters in anthropogenic systems are 'natural', or accidents resulting from external disruptions. Another misconception is that our complex systems could be well controlled or that our socio-economic system would automatically fix itself.

Such ways of thinking impose huge risks on society. However, owing to the systemic nature of man-made disasters, it is hard to blame anybody for the damage. Therefore, classical self-adjustment and feedback mechanisms will not ensure responsible action to avert possible disasters. It also seems that present law cannot handle situations well, when the problem does not lie in the behaviour of individuals or companies, but in the interdependencies between them.

The increasing availability of 'big data' has raised the expectation that we could make the world more predictable and controllable. Indeed, real-time management may overcome instabilities caused by delayed feedback or lack of information. However, there are important limitations: too much data can make it difficult to separate reliable from ambiguous or incorrect information, leading to misinformed decision-making. Hence too much information may create a more opaque rather than a more transparent picture.

If a country had all the computer power in the world and all the data, would this allow a government to make the best decisions for everybody? Not necessarily. The principle of a caring state (or benevolent dictator) would not work, because the world is too complex to be optimized top-down in real time. Decentralized coordination with affected (neighbouring) system components can achieve better results, adapted to local needs[45]. This means that a participatory approach, making use of local resources, can be more successful. Such an approach is also more resilient to perturbations.

For today's anthropogenic system, predictions seem possible only over short time periods and in a probabilistic sense. Having all the data in the world would not allow one to forecast the future. Nevertheless, one can determine under what conditions systems are prone to cascades or not. Moreover, weak system components can be used to produce early warning signals. If safety precautions are lacking, however, spontaneous cascades might be unstoppable and become catastrophic. In other words, predictability and controllability are a matter of proper systems design and operation. It will be a twentyfirst-century challenge to learn how to turn this into practical solutions and how to use the positive sides of cascade effects. For example, cascades can produce a large-scale coordination of traffic lights[45] and vehicle flows[70], or promote the spreading of information and innovations[57,58], of happiness[71], social norms[72], and cooperation[31,32,59]. Taming cascade effects could even help to mobilize the collective effort needed to address the challenges of the century ahead.

Received 31 August 2012; accepted 26 February 2013.

1. World Economic Forum. *Global Risks* 2012 and *2013* (WEF, 2012 and 2013); http://www.weforum.org/issues/global-risks.
2. Rinaldi, S. M., Peerenboom, J. P. & Kelly, T. K. Critical infrastructure interdependencies. *IEEE Control Syst.* **21,** 11–25 (2001).
3. Rosato, V. *et al.* Modelling interdependent infrastructures using interacting dynamical models. *Int. J. Critical Infrastruct.* **4,** 63–79 (2008).
4. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature* **464,** 1025–1028 (2010).
5. Gao, J., Buldyrev, S. V., Havlin, S. & Stanley, H. E. Robustness of networks of networks. *Phys. Rev. Lett.* **107,** 195701 (2011).
6. Vespignani, A. The fragility of interdependency. *Nature* **464,** 984–985 (2010).
7. Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439,** 462–465 (2006).
8. Vespignani, A. Predicting the behavior of techno-social systems. *Science* **325,** 425–428 (2009).
9. Epstein, J. M. Modelling to contain pandemics. *Nature* **460,** 687 (2009).
10. Crutzen, P. & Stoermer, E. The anthropocene. *Global Change Newsl.* **41,** 17–18 (2000).

11. Helbing, D. & Carbone, A. (eds) Participatory science and computing for our complex world. *Eur. Phys. J. Spec. Top.* **214,** (special issue) 1–666 (2012).
12. Zeeman, E. C. (ed.) *Catastrophe Theory* (Addison-Wesley, 1977).
13. Stanley, H. E. *Introduction to Phase Transitions and Critical Phenomena* (Oxford Univ. Press, 1987).
14. Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99,** 5766–5771 (2002).
15. Motter, A. E. Cascade control and defense in complex networks. *Phys. Rev. Lett.* **93,** 098701 (2004).
16. Simonsen, I., Buzna, L., Peters, K., Bornholdt, S. & Helbing, D. Transient dynamics increasing network vulnerability to cascading failures. *Phys. Rev. Lett.* **100,** 218701 (2008).
17. Little, R. G. Controlling cascading failure: understanding the vulnerabilities of interconnected infrastructures. *J. Urban Technol.* **9,** 109–123 (2002).
This is an excellent analysis of the role of interconnectivity in catastrophic failures.
18. Buzna, L., Peters, K., Ammoser, H., Kühnert, C. & Helbing, D. Efficient response to cascading disaster spreading. *Phys. Rev. E* **75,** 056107 (2007).
19. Lorenz, J., Battiston, S. & Schweitzer, F. Systemic risk in a unifying framework for cascading processes on networks. *Eur. Phys. J. B* **71,** 441–460 (2009).
This paper gives a good overview of different classes of cascade effects with a unifying theoretical framework.
20. Battiston, S., Delli Gatti, D., Gallegati, M., Greenwald, B. & Stiglitz, J. E. Default cascades: when does risk diversification increase stability? *J. Financ. Stab.* **8,** 138–149 (2012).
21. Albeverio, S., Jentsch, V. & Kantz, H. (eds) *Extreme Events in Nature and Society* (Springer, 2010).
22. Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality: an explanation of the 1/*f* noise. *Phys. Rev. Lett.* **59,** 381–384 (1987).
23. Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* **406,** 378–382 (2000).
24. Kun, F., Carmona, H. A., Andrade, J. S. Jr & Herrmann, H. J. Universality behind Basquin's law of fatigue. *Phys. Rev. Lett.* **100,** 094301 (2008).
25. Achlioptas, D., D'Souza, R. M. & Spencer, J. Explosive percolation in random networks. *Science* **323,** 1453–1455 (2009).
26. Sornette, D. & Ouillon, G. Dragon-kings: mechanisms, statistical methods and empirical evidence. *Eur. Phys. J. Spec. Top.* **205,** 1–26 (2012).
27. Nicolis, G. *Introduction to Nonlinear Science* (Cambridge Univ. Press, 1995).
28. Strogatz, S. H. *Nonlinear Dynamics and Chaos* (Perseus, 1994).
29. Liu, Y. Y., Slotine, J. J. & Barabasi, A. L. Controllability of complex networks. *Nature* **473,** 167–173 (2011).
30. Dörner, D. *The Logic of Failure* (Metropolitan, 1996).
This book is a good demonstration that we tend to make wrong decisions when trying to manage complex systems.
31. Nowak, M. A. *Evolutionary Dynamics* (Belknap, 2006).
32. Helbing, D. *Social Self-Organization* (Springer, 2012).
This book offers an integrative approach to agent-based modelling of emergent social phenomena, systemic risks in social and economic systems, and how to manage complexity.
33. Johansson, A., Helbing, D., Al-Abideen, H. Z. & Al-Bosta, S. From crowd dynamics to crowd safety: a video-based analysis. *Adv. Complex Syst.* **11,** 497–527 (2008).
34. Helbing, D. & Mukerji, P. Crowd disasters as systemic failures: analysis of the Love Parade disaster. *Eur. Phys. J. Data Sci.* **1,** 7 (2012).
35. Bettencourt, L. M. A. *et al.* Growth, innovation, scaling and the pace of life in cities. *Proc. Natl Acad. Sci. USA* **104,** 7301–7306 (2007).
36. Ball, P. *Why Society is a Complex Matter* (Springer, 2012).
37. Aven, T. & Vinnem, J. E. (eds) *Risk, Reliability and Societal Safety* Vols 1–3 (Taylor and Francis, 2007).
This compendium is a comprehensive source of information about risk, reliability, safety and resilience.
38. Rodriguez, H., Quarantelli, E. L. & Dynes, R. R. (eds) *Handbook of Disaster Research* (Springer, 2007).
39. Cox, L. A. Jr. *Risk Analysis of Complex and Uncertain Systems* (Springer, 2009).
40. Perrow, C. *Normal Accidents. Living with High-Risk Technologies* (Princeton Univ. Press, 1999).
This eye-opening book shows how catastrophes result from couplings and complexity.
41. Peters, G. A. & Peters, B. J. *Human Error. Causes and Control* (Taylor and Francis, 2006).
This book is a good summary of why, how and when people make mistakes.
42. Clarke, L. *Worst Cases* (Univ. Chicago, 2006).
43. Axelrod, R. & Cohen, M. D. *Harnessing Complexity* (Basis Books, 2000).
This book offers a good introduction into complex social systems and bottom-up management.
44. Tumer, K. & Wolpert, D. H. *Collectives and the Design of Complex Systems* (Springer, 2004).
45. Lämmer, S. & Helbing, D. Self-control of traffic lights and vehicle flows in urban road networks. *J. Stat. Mech.* P04019 (2008).
46. Perkins, C. E. & Royer, E. M. Ad-hoc on-demand distance vector routing. In *Second IEEE Workshop on Mobile Computing Systems and Applications* 90–100 (WMCSA Proceedings, 1999).
47. Amin, M. M. & Wollenberg, B. F. Toward a smart grid: power delivery for the 21st century. *IEEE Power Energy Mag.* **3,** 34–41 (2005).
48. Schneider, C. M., Moreira, A. A., Andrade, J. S. Jr, Havlin, S. & Herrmann, H. J. Mitigation of malicious attacks on networks. *Proc. Natl Acad. Sci. USA* **108,** 3838–3841 (2011).
49. Comfort, L. K., Boin, A. & Demchak, C. C. (eds) *Designing Resilience. Preparing for Extreme Events* (Univ. Pittsburgh, 2010).
50. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461,** 53–59 (2009).
51. Pikovsky, A., Rosenblum, M. & Kurths, J. *Synchronization* (Cambridge Univ. Press, 2003).
52. Haldane, A. G. & May, R. M. Systemic risk in banking ecosystems. *Nature* **469,** 351–355 (2011).
53. Battiston, S., Puliga, M., Kaushik, R., Tasca, P. & Caldarelli, G. DebtRank: too connected to fail? Financial networks, the FED and systemic risks. *Sci. Rep.* **2,** 541 (2012).
54. Stiglitz, J. E. *Freefall: America, Free Markets, and the Sinking of the World Economy* (Norton & Company, 2010).
55. Sterman, J. *Business Dynamics: Systems Thinking and Modeling for a Complex World* (McGraw-Hill/Irwin, 2000).
56. Helbing, D. & Lämmer, S. in *Networks of Interacting Machines: Production Organization in Complex Industrial Systems and Biological Cells* (eds Armbruster, D., Mikhailov, A. S. & Kaneko, K.) 33–66 (World Scientific, 2005).
57. Young, H. P. Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. *Am. Econ. Rev.* **99,** 1899–1924 (2009).
58. Montanari, A. & Saberi, A. The spread of innovations in social networks. *Proc. Natl Acad. Sci. USA* **107,** 20196–20201 (2010).
59. Grund, T., Waloszek, C. & Helbing, D. How natural selection can create both self- and other-regarding preferences, and networked minds. *Sci. Rep.* **72,** 1480, http://dx.doi.org/10.1038/srep01480 (2013).
60. Lazer, D. *et al.* Computational social science. *Science* **323,** 721–723 (2009).
61. Epstein, J. M. & Axtell, R. L. *Growing Artificial Societies: Social Science from the Bottom Up* (Brookings Institution, 1996).
This is a groundbreaking book on agent-based modelling.
62. Gilbert, N. & Bankes, S. Platforms and methods for agent-based modeling. *Proc. Natl Acad. Sci. USA* **99** (S3), 7197–7198 (2002).
63. Farmer, J. D. & Foley, D. The economy needs agent-based modeling. *Nature* **460,** 685–686 (2009).
64. Szell, M., Sinatra, R., Petri, G., Thurner, S. & Latora, V. Understanding mobility in a social petri dish. *Sci. Rep.* **2,** 457 (2012).
65. de Freitas, S. Game for change. *Nature* **470,** 330–331 (2011).
66. McNeil, A. J., Frey, R. & Embrechts, P. *Quantitative Risk Management* (Princeton Univ. Press, 2005).
67. Preis, T., Kenett, D. Y., Stanley, H. E., Helbing, D. & Ben-Jacob, E. Quantifying the behaviour of stock correlations under market stress. *Sci. Rep.* **2,** 752 (2012).
68. Floriano, D. & Mattiussi, C. *Bio-Inspired Artificial Intelligence* (MIT Press, 2008).
69. Pentland, A. Society's nervous system: building effective government, energy, and public health systems. *IEEE Computer* **45,** 31–38 (2012).
70. Kesting, A., Treiber, M., Schönhof, M. & Helbing, D. Adaptive cruise control design for active congestion avoidance. *Transp. Res. C* **16,** 668–683 (2008).
71. Fowler, J. H. & Christakis, N. A. Dynamic spread of happiness in a large social network. *Br. Med. J.* **337,** a2338 (2008).
72. Helbing, D. & Johansson, A. Cooperation, norms, and revolutions: a unified game-theoretical approach. *PLoS ONE* **5,** e12530 (2010).
73. Seydel, R. U. *Practical Bifurcation and Stability Analysis* (Springer, 2009).
74. Bak, P., Christensen, K., Danon, L. & Scanlon, T. Unified scaling law for earthquakes. *Phys. Rev. Lett.* **88,** 178501 (2002).
75. Helbing, D. Traffic and related self-driven many-particle systems. *Rev. Mod. Phys.* **73,** 1067–1141 (2001).
76. Lozano, S., Buzna, L. & Diaz-Guilera, A. Role of network topology in the synchronization of power systems. *Eur. Phys. J. B* **85,** 231–238 (2012).
77. Schuster, H. G. & Just, W. *Deterministic Chaos* (Wiley-VCH, 2005).
78. Wiener, N. *Cybernetics* (MIT Press, 1965).
79. Beale, N. *et al.* Individual versus systemic risk and the regulator's dilemma. *Proc. Natl Acad. Sci. USA* **108,** 12647–12652 (2011).
80. Allen, P. M. Evolution, population dynamics, and stability. *Proc. Natl Acad. Sci. USA* **73,** 665–668 (1976).
81. Tainter, J. *The Collapse of Complex Societies* (Cambridge Univ. Press, 1988).
82. The World Economic Forum, *Global Risks 2011* 6th edn (WEF, 2011); http://reports.weforum.org/wp-content/blogs.dir/1/mp/uploads/pages/files/global-risks-2011.pdf.
83. Huntington, S. P. The clash of civilisations? *Foreign Aff.* **72,** 22–49 (1993).
84. Cederman, L. E. Endogenizing geopolitical boundaries with agent-based modeling. *Proc. Natl Acad. Sci. USA* **99** (suppl. 3), 7296–7303 (2002).
85. Johnson, N. *et al.* Pattern in escalations in insurgent and terrorist activity. *Science* **333,** 81–84 (2011).
86. Beck, U. *Risk Society* (Sage, 1992).
87. Lin, N. *Social Capital* (Routeledge, 2010).
88. Kröger, W. & Zio, E. *Vulnerable Systems* (Springer, 2011).

Acknowledgements This work has been supported partially by the FET Flagship Pilot Project FuturICT (grant number 284709) and the ETH project "Systemic Risks—Systemic Solutions" (CHIRP II project ETH 48 12-1). I thank L. Böttcher, T. Grund, M. Kaninia, S. Rustler and C. Waloszek for producing the cascade spreading movies and figures. I also thank the FuturICT community for many inspiring discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.H. (dhelbing@ethz.ch).

doi:10.1038/nature11909

Using membrane transporters to improve crops for sustainable food production

Julian I. Schroeder[1], Emmanuel Delhaize[2], Wolf B. Frommer[3], Mary Lou Guerinot[4], Maria J. Harrison[5], Luis Herrera-Estrella[6], Tomoaki Horie[7], Leon V. Kochian[8], Rana Munns[2,9], Naoko K. Nishizawa[10], Yi-Fang Tsay[11] & Dale Sanders[12]

With the global population predicted to grow by at least 25 per cent by 2050, the need for sustainable production of nutritious foods is critical for human and environmental health. Recent advances show that specialized plant membrane transporters can be used to enhance yields of staple crops, increase nutrient content and increase resistance to key stresses, including salinity, pathogens and aluminium toxicity, which in turn could expand available arable land.

Of the present global population of seven billion people, almost one billion are undernourished and lack sufficient protein, fats and carbohydrates in their diets[1]. An additional billion people are malnourished because their diets lack required micronutrients such as iron, zinc and vitamin A (ref. 2). These dietary deficiencies have an enormous negative impact on global health, resulting in increased susceptibility to infection and diseases, as well as increasing the risk of significant mental impairment[3]. During the next four decades, an expected additional two billion humans will require nutritious food. Along with growing urbanization, increased demand for protein in developing countries, coupled with impending climate change and population growth, will impose further pressures on agricultural production[4]. Global demand for food is predicted to increase by 40% by 2030 (ref. 4). Innovative solutions are required to increase production on the land currently used for agriculture, because we are already close to the sustainable limit of 15% of the Earth's surface that can be exploited for crop production[5].

Analysis of crop yields globally shows that in those developing regions where humans are most susceptible to malnutrition, the availabilities of inorganic nutrients and water are the principal factors that determine crop productivity[6,7]. Simply increasing inorganic fertilizer use and water supply or applying organic farming systems to agriculture[8] will be unable to satisfy the joint requirements of increased yield and environmental sustainability. Increasing food production on limited land resources will rely on innovative agronomic practices coupled to the genetic improvement of crops[9].

Transport proteins embedded within membranes are key targets for improving the efficiency with which plants take up and use water and nutrients. These proteins not only transport mineral nutrients and control drought tolerance but are also essential for moving sucrose, the energy currency of plants, to where it is needed. Furthermore, transporters are also central to mechanisms that allow plants to tolerate adverse environments such as saline or acid soils. Advances driven by physiology, genetics and biophysics over the past 20 years have dramatically improved our understanding of the molecular basis of plant nutrition and how plants respond to stress. Genome sequencing and the development of experimental systems for studying transporter function have allowed many of the major families of membrane transporters to be characterized. Next-generation sequencing is leading to an understanding of how the natural genetic diversity of plant membrane transporters can be exploited for agriculture, whether by marker-assisted breeding or through genetic engineering. Breakthrough approaches involving transcriptional-activator-like effectors and genome editing[10] can provide non-genetically modified (non-GM) yet molecularly directed rapid solutions to crop improvement.

Here we report on findings demonstrating that understanding the biology of plant membrane transporters can be a key contributor to the goal of global food security. We discuss examples where fundamental research is already being translated into practical applications such as enhancing the micronutrient content of grain and improving the plant tolerance to saline and acidic soils. We further discuss potential applications linked to breakthroughs in basic research that are yet to be applied to crop plants. This Perspective reviews the extent to which the rapid advances in plant transport research address global aspects of food security, and how we can potentially reduce the time between trait identification in the laboratory and exploitation in the field.

Transporters, stress resistance and yield

Aluminium-tolerant crops for acid soils

Acid soils comprise 30% of Earth's ice-free land and thus constrain agricultural production, given that only a small proportion of these soils is suitable for crops[11]. At soil pH values above 5, aluminium exists in the soil in non-toxic complexed forms. However, when soils are acidic, Al^{3+} ions are freed in the soil, resulting in plant toxicity. Once in the soil solution, Al^{3+} damages the root tips of susceptible plants and inhibits root growth, which impairs the uptake of water and nutrients. Natural genetic variation in Al^{3+} tolerance exists within major cereal crops. The efflux of organic anions from roots was discovered to be a naturally occurring tolerance mechanism of several species[12]. Transport proteins are central to this mechanism, with members of two families of transport proteins responsible for exporting the organic anions from inside root cells to the external medium surrounding roots. The organic anions secreted by roots chelate Al^{3+} into a non-toxic form, thus protecting the sensitive tips and allowing the roots to grow unimpeded (Fig. 1).

In wheat, the *Triticum aestivum* aluminium-activated malate transporter 1 gene *TaALMT1* encodes an Al^{3+}-gated anion channel that

[1]Division of Biological Sciences, Food and Fuel for the 21st Century Center, University of California San Diego, La Jolla, California 92093-0116, USA. [2]CSIRO Plant Industry, Canberra 2601, Australia. [3]Department of Plant Biology, Carnegie Institution for Science, Stanford, California 94305, USA. [4]Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire 03755, USA. [5]Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, New York 14853, USA. [6]Laboratorio Nacional de Genomica para la Biodiversidad, Centro de Investigación y de Estudios Avanzados, 36500 Irapuato, Mexico. [7]Division of Applied Biology, Faculty of Textile Science and Technology, Shinshu University, Nagano 386-8567, Japan. [8]Robert Holley Center for Agriculture and Health, USDA-ARS, Cornell University, Ithaca, New York 14853, USA. [9]School of Plant Biology, University of Western Australia, Crawley 6009, Australia. [10]Research Institute for Bioresources and Biotechnology, Ishikawa Prefectural University, Ishikawa 921-8836, Japan. [11]Institute of Molecular Biology, Academia Sinica, Taipei 11529, Taiwan. [12]John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK.

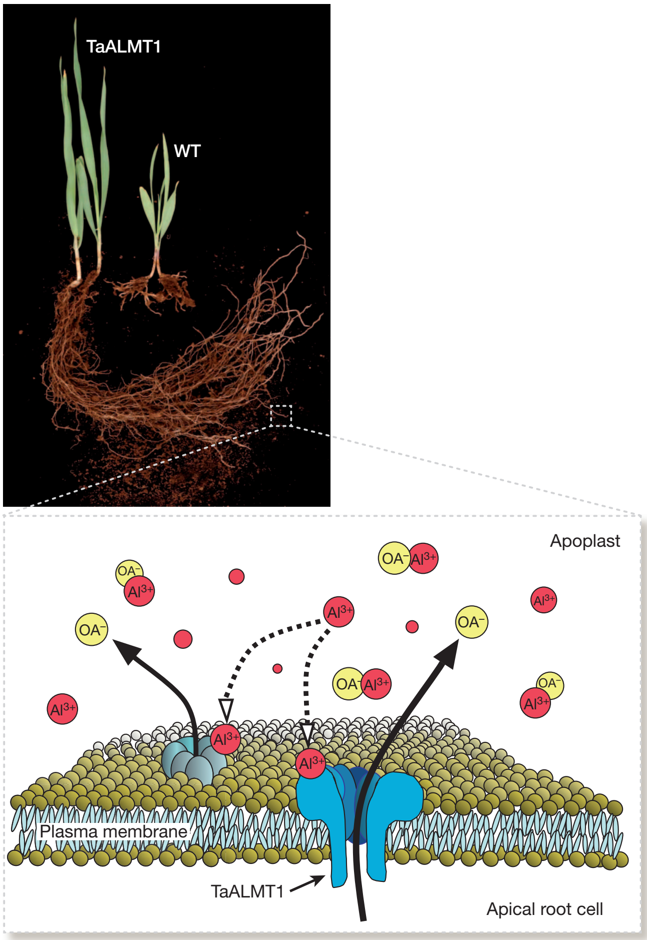


Figure 1 | Engineering plants for enhanced aluminium (Al^{3+}) tolerance. The photograph shows barley seedlings grown on an acid soil that contains high concentrations of toxic Al^{3+}. One seedling has been genetically engineered with an Al^{3+}-tolerance transporter gene from wheat (*TaALMT1*), whereas the other seedling is the non-transgenic parental line (wild type, WT). The diagram shows the TaALMT1 anion channel (blue structure) embedded within the plasma membrane of apical root cells. In acid soils, Al^{3+} activates TaALMT1 (dashed line) resulting in malate efflux into the apoplast (cell wall) external to the cytoplasm. Malate molecules (OA^{-}, yellow circles) bind Al^{3+} in the apoplast to protect cells from aluminium toxicity at the root apex. The diagram is modified from figure 2 in ref. 92.

facilitates malate efflux from roots[13] (Fig. 1). Molecular markers based on the *TaALMT1* gene can be used in marker-assisted breeding to select for Al^{3+} tolerance in wheat germplasm. The *TaALMT1* gene itself can be used to genetically modify susceptible species for improved Al^{3+} tolerance. When expressed in barley, one of the most Al^{3+}-sensitive cereal crops, *TaALMT1* confers substantially improved grain yields in acid soil[14] (Fig. 1). Similarly, a unique subgroup of the large family of plant multidrug and toxic compound extrusion (MATE) transporters mediates citrate efflux from root cells[15,16]. In sorghum[16], barley[17], and maize[18], MATE transporters located at the root tip confer Al^{3+}-activated citrate efflux and represent the primary Al^{3+} tolerance proteins. Genetic studies in sorghum recently identified markers associated with Al^{3+}-tolerant alleles of the *Sorghum bicolor* MATE gene *SbMATE*. These markers have been used by breeders to introgress rapidly the most favourable *SbMATE* alleles into sorghum germplasm, which is currently being field-tested in acid soils[16].

Although genes encoding organic anion transporters are already established as a means of enhancing the Al^{3+} tolerance of crops, other recently discovered mechanisms provide additional options. For instance, rice is the most Al^{3+} tolerant of the cereal crops and uses mechanisms distinct from the efflux of organic anions. One of these mechanisms involves transporters working in concert ultimately to sequester Al^{3+} into the vacuole, thus removing Al^{3+} from mainstream metabolism[19,20]. Genetic analysis has identified susceptible and tolerant variants of one of the transporter genes (natural resistance-associated macrophage protein aluminium transporter 1, *Nrat1*) that explains a large proportion of the variation in Al^{3+} tolerance within the rice *aus* subpopulation. This finding provides a promising tool for marker-assisted breeding[21].

The discovery of transporters that mediate Al^{3+} tolerance has identified two principal strategies that plants use to deal with this toxic cation. In one strategy, Al^{3+} is excluded from cells by chelating the toxic ion external to plants and in the other Al^{3+} is sequestered within cells in the vacuole. The genes encoding these transporters can be used to develop Al^{3+}-tolerant crops and represent an important component—along with management practices such as soil liming to increase soil pH—of a strategy for improving yields on acid soils.

HKT transporters improve salt tolerance

Approximately 7% of the world's land including agricultural lands is affected by either salinity or sodium toxicity. Production in over 30% of irrigated crops and 7% of dryland agriculture worldwide is limited by salinity stress. Crop irrigation is increasing soil salinity, owing to trace amounts of salt in irrigation waters. Plant plasma membrane transporters in the HKT family transport sodium (Na^{+}) and potassium (K^{+}) (ref. 22) and play an essential part in salt tolerance[23]. Research in the reference plant *Arabidopsis* showed that the 'class 1' HKT transporters are Na^{+} selective and protect plant leaves from salinity stress by prohibiting toxic sodium over-accumulation in leaves[23]. Class 1 HKT transporters are expressed in veins[23] that connect nutrient flux between roots and leaves. These transporters are expressed in the living cells surrounding the xylem, which are vessels that carry nutrients and water to the leaves. Class 1 HKT transporters remove excess Na^{+} from the xylem in *Arabidopsis* and rice, thereby keeping Na^{+} below toxic levels in the photosynthetic leaf tissues[24–26] (Fig. 2). Analogous mechanisms have

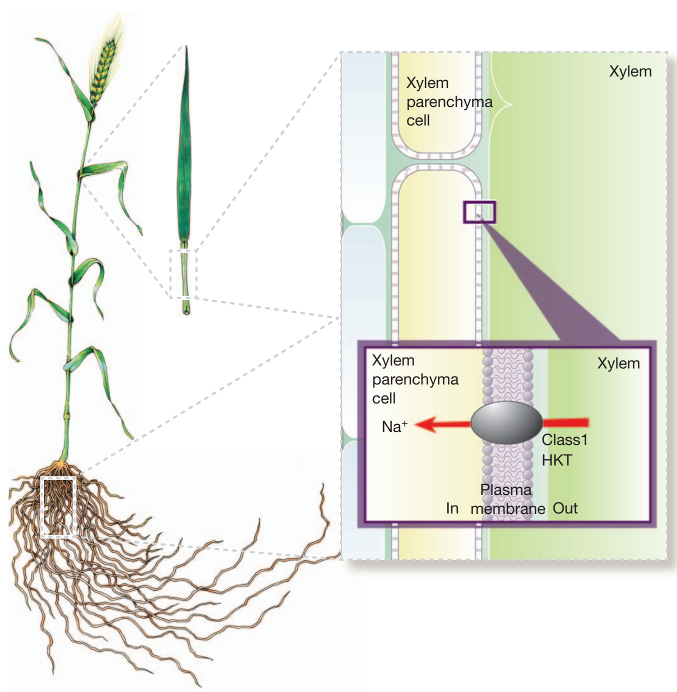


Figure 2 | HKT transporter-mediated salt tolerance in plants. The drawing illustrates the function of class I HKT transporters in protecting plants from salinity stress. These HKT transporters mediate Na^{+} unloading from the xylem under salinity stress, which prevents Na^{+} over-accumulation in leaves, thereby protecting photosynthetic organs. An example of this mechanism in wheat plants is shown.

been demonstrated in wheat for the *HKT1;4* and *HKT1;5* genes[27,28] (Fig. 2). Remarkably, the recent introgression of an ancestral form of the *HKT1;5* gene from the more Na^+-tolerant wheat relative *Triticum monococcum* into susceptible commercial durum wheat (*Triticum turgidum* ssp *durum*) increased grain yields on saline soil by 25% in the field, illustrating the immense potential of this mechanism[28].

Some crops are salt tolerant through the effective sequestration of Na^+ in leaf vacuoles by Na^+/H^+ antiporters[29]. Specific 'class 2' HKT transporters[30] mediate cation influx into roots[31]. These class 2 HKT transporters, together with transporters that sequester sodium and potassium in the vacuole[32,33], have the potential to improve the production of cereals such as barley, a species that copes with high Na^+ loads in leaves by compartmentation in the vacuole[34]. Thus combining (pyramiding) HKT transporter traits with vacuolar Na^+ sequestration mechanisms provides a potentially powerful platform for molecular breeding and transgenic approaches to improve the salinity tolerance of crops.

SWEET transporters and pathogen resistance

Photosynthesis in leaves produces sugars, which are distributed through the veins to support growth of roots, meristems and seeds. The translocation rates and relative distribution critically determine the yield potential of crops. Sucrose is a key energy-carrying molecule that also acts as the driver for translocation of all other nutrients and signalling molecules in veins, but until recently, the molecular players were unknown. In the early 1990s the sucrose-proton co-transporter SUT1 was identified as the key transporter that loads sugar into veins[35], yet the mechanism for sucrose release from the leaf cells that synthesize it remained elusive. Recently, the 'SWEET' sugar transporters were identified with the help of sugar sensors based on Förster (fluorescence) resonance energy transfer technology[36,37]. Known crop genomes possess about 20 SWEET genes. SWEETs are plasma membrane proteins located in the phloem parenchyma, a cell type inside the veins that exports sucrose to the SUT1 sugar loaders (Fig. 3). The import and export of sucrose from vein cells is controlled by hormones, turgor feedback and sugar levels[38]. Knowledge of this machinery could provide a new starting point to engineer yield by modifying energy and carbon distribution within the plant.

Notably, SWEET sugar efflux transporters have been identified as pathogen resistance loci, leading to a new understanding of disease development in plants[39–42]. The growth of pathogens in leaves and stems depends on nutrient supply from their plant hosts. Blight bacteria directly induce SWEET gene expression in rice in infected cells through transcriptional-activator-like effectors (bacterial transcription factors that directly target SWEET promoters). Inhibiting the induction of SWEET genes with an innovative technology such as chromosomal editing of the promoters of SWEETs with TALENs (artificial transcriptional-activator-like effector nucleases) or through cell-specific expression of microRNAs in cells outside the phloem has now enabled blight resistance to be engineered in rice[43,44]. The discovery of these key players in combination with TALENs promises new ways of engineering both crop yield and pathogen resistance, without the introduction of foreign genetic material, to produce plants with significantly improved performance in the field.

Human health and plant nutrition

Pumping iron and zinc

Over two billion people suffer from iron and zinc deficiencies because their plant-based diets are not a sufficiently rich source of these essential elements. Developing crop cultivars with increased micronutrient concentrations, an approach known as biofortification, is challenging because metal ion concentrations in various tissues and compartments are maintained within narrow physiological limits by coordinated uptake, translocation and storage. Furthermore, for crops like rice, removal of the outer layers of the grain during polishing essentially removes all of the micronutrients, leaving only the starchy endosperm.

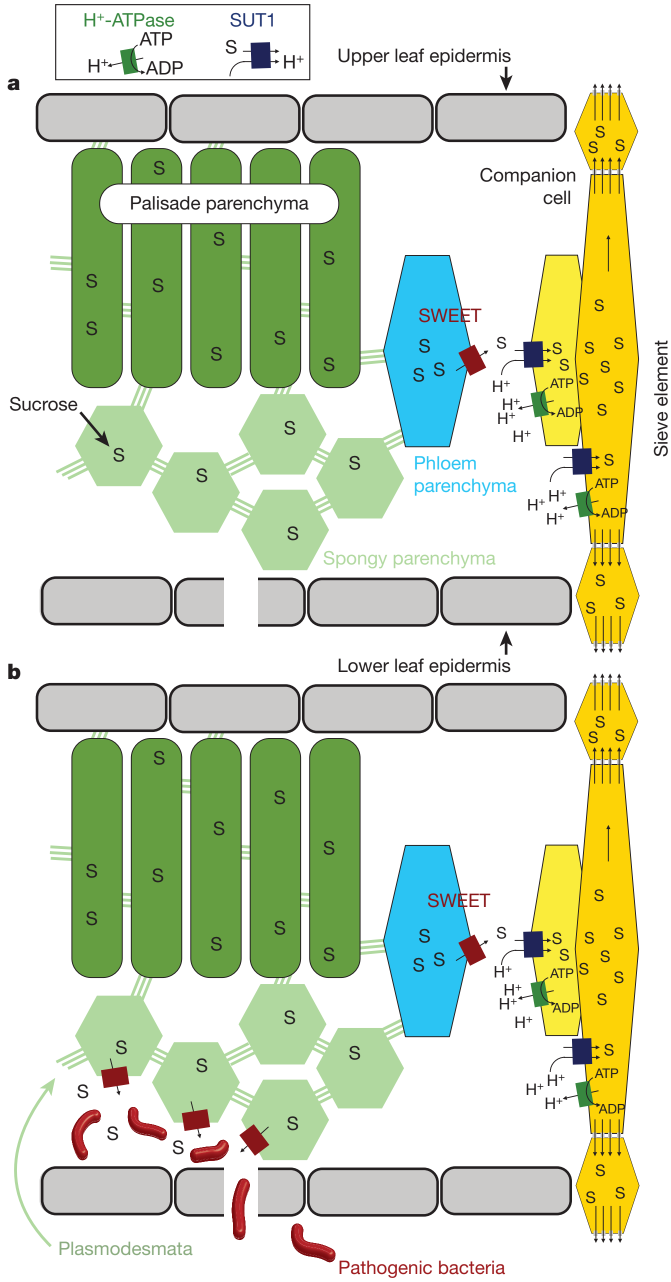


Figure 3 | The role of SWEET sugar transporters in efflux of sucrose into the cell-wall space and induction by pathogenic bacteria. a, SWEETs (red) localized in the phloem parenchyma (a cell type of the plant vasculature), export sucrose produced by photosynthesis into the cell wall, from where it is loaded actively, with the help of the transporter SUT1 and energized by H^+-ATPases into the actual conduits, the sieve element companion cell complex for translocation to seeds. Photosynthesis mainly occurs in the palisade parenchyma. **b**, The role of SWEETs as the 'Achilles' heel' (susceptibility factors) of host plants during pathogen infection. SWEETs are induced directly as a consequence of the injection of transcriptional-activator-like effectors from pathogens via type III secretion systems into the infected plant cell, leading to release of sugars as a critical source of nutrition for the pathogens.

By expressing key genes involved in the mobilization of micronutrients from the soil to the seed, scientists have biofortified rice, a staple food consumed by half the world population every day (Fig. 4). Enhancing iron translocation through overproduction of the metal chelator nicotianamine and phytosiderophores[45–47] or enhancing iron influx into the endosperm by means of the iron-nicotianamine transporter *Oryza sativa* yellow-stripe-like-2 (OsYSL2)[48], has resulted in greenhouse-grown rice with three- to fourfold higher levels of iron (Fe) in polished grain.

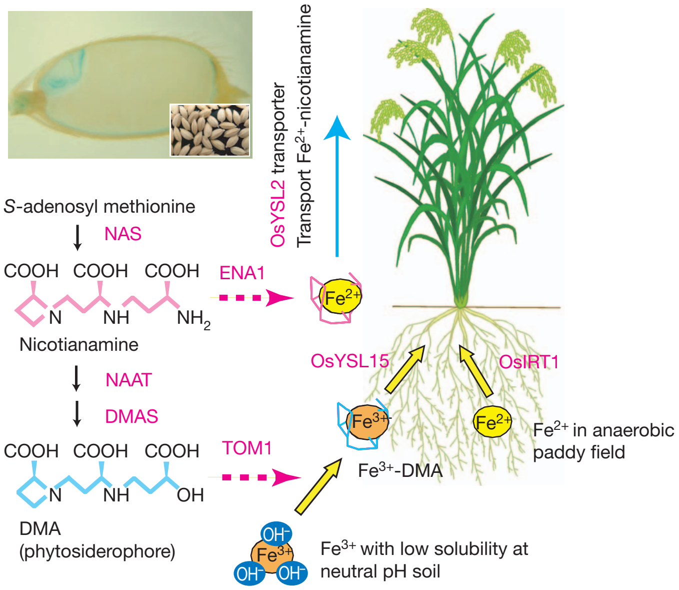


Figure 4 | Iron transport in rice. Rice takes up iron from the soil as Fe^{3+} deoxymugineic acid (DMA) by the OsYSL15 transporter. Rice also uses the OsIRT1 transporter to take up Fe^{2+}, which is abundant in submerged and anaerobic conditions. DMA, which is the primary phytosiderophore that aids in iron transport, is synthesized from *S*-adenosyl methionine through three sequential enzymatic reactions mediated by nicotianamine synthase (NAS), nicotianamine aminotransferase (NAAT), and DMA synthase (DMAS), and then secreted by the efflux transporter TOM1 to solubilize iron in the soil. Nicotianamine, which is the biosynthetic precursor of DMA, is a chelator of divalent metals and plays a part in translocation of metals within plants. Nicotianamine is secreted into the cell wall by the nicotianamine efflux transporter ENA1. The iron–nicotianamine transporter OsYSL2 mediates iron influx into rice grains. The photograph shows iron staining (blue coloration) of a rice seed (inset shows rice seeds). Iron is mainly localized to the embryo and the outer layers of the grain.

Combining overproduction of nicotianamine with enhanced expression of the iron storage protein ferritin increased the iron content more than sixfold[49], and combining all three approaches has resulted in paddy-field-grown polished rice with Fe concentrations 4.4-fold higher than those found in non-transgenic seeds, with no yield penalty[50]. Although these results are impressive and bring iron levels close to those recommended by nutritionists, only a handful of studies have tested whether these enhanced levels of nutrients are bioavailable. Most encouragingly, enhancing the nicotianamine concentration does increase the levels of bioavailable iron[46] and zinc[51] in polished rice.

Vacuolar sequestration is another mechanism to enhance the concentrations of iron and zinc (Zn) in seeds[52]. Transporters belonging to several different families transport metals between the cytoplasm and the vacuole[53–55], including the *Arabidopsis* vacuolar iron transporter *VIT1* protein, which is highly expressed in developing seeds and transports iron and manganese into the vacuole[54]. Disruption of the rice *VIT* orthologues (*OsVIT1* and *OsVIT2*) increases Fe/Zn accumulation in rice seeds and decreases Fe/Zn in the source organ flag leaves, probably because *VIT* genes are highly expressed in rice flag leaves. Without a sink, there is enhanced Fe/Zn translocation to the seed, providing another strategy with which to biofortify Fe/Zn in staple foods[56]. Metal tolerance protein 1 (MTP1) also transports divalent cations into the vacuole and is another promising candidate for use in biofortification[57]. Thus several strategies are being used to enhance iron and zinc micronutrients in edible plant tissues, but more improvements are needed. We can use our growing knowledge of the transporters that take up micronutrients from the soil, such as iron-regulated transporter 1 (IRT1)[58], the major entry point for Fe in many plant species. Enhanced nutrient content is a crucial goal in the light of the world's growing population and the central roles of staple crops in human diets.

Enhancing phosphate use efficiency

Phosphorus (P) is a macro-element that is essential for plant growth and of vital importance to crop yield. The availability of inorganic P, or orthophosphate (the only form of P directly accessible to plants), is influenced by the biogeochemical properties of the soil and limits crop production on nearly 70% of the world's agricultural soils[59]. Consequently, global crop production depends on orthophosphate fertilizers, which are produced from rock phosphate, a finite, non-renewable mineral resource (Fig. 5a). Only 20–30% of the P fertilizer applied is used by cultivated plants and at the current rate of use, it is estimated that rock phosphate reserves will be consumed within the next 70–200 years[60], so ensuring the sustainable use of orthophosphate is of paramount importance for human nutrition. Improving orthophosphate acquisition and use-efficiency in plants is a complex problem and recent solutions have included modifications to root growth and architecture[61,62], and novel engineering strategies to use alternative P sources[60]. An understanding of plant transporter proteins may offer additional approaches. Plants possess several families of orthophosphate transporter proteins, and both high- and low-affinity transporters are important for orthophosphate uptake into roots[63,64]. Phosphate transporters are also critical for orthophosphate distribution throughout the plant, and for remobilization between source and sink tissues[65,66]. A phosphate efflux transporter (PHO1), essential for orthophosphate transfer to the shoot, is a major player in the regulation of orthophosphate homeostasis[67] and may provide strategies for optimizing orthophosphate distribution within plants.

In addition to the direct acquisition of orthophosphate from soil, most crop species have the capacity to form symbiotic associations (called arbuscular mycorrhizae) with soil fungi. These fungi capture orthophosphate through extensive hyphal networks and deliver it to symbiotic compartments in the root, where plant orthophosphate transporters transfer this ion into the root cells. Plant symbiotic orthophosphate transporter function is essential for this process and is also required to maintain the symbiosis[68,69]. Phosphate transporters are important targets for breeding plants with improved orthophosphate acquisition and use-efficiency, and that benefit maximally from their fungal symbionts.

Nitrate sensing and transport

The application of nitrogen (N) fertilizers has greatly increased crop yields and alleviated hunger over the past five decades. However, N fertilizer production consumes 1% of global energy usage and poses the highest input cost for many crops. Nevertheless, only 30% to 50% of the N fertilizer applied is used by plants. The remainder can lead to production of the greenhouse gas N_2O, or to eutrophication of aquatic ecosystems through water run-off. Therefore, enhancing crop nitrogen utilization efficiency is an important goal[70]. For most crops, nitrate is the primary nitrogen source and so enhancing nitrate uptake is one strategy for improving nitrogen utilization efficiency.

Multiple nitrate uptake transporters of the NRT1 and NRT2 families work together to enable nitrogen uptake in plants[71,72]. Most NRT1 transporters are low-affinity nitrate transporters, meaning that they function mainly when nitrate is abundant. In contrast, chlorate resistant 1 (CHL1, also known as NRT1.1) is a dual-affinity nitrate transporter involved in nitrate acquisition[73,74]. A recent study found that CHL1 also functions as a nitrate sensor, thus regulating nitrate-induced gene expression[75], which implies that plants use this transporter to monitor changes in external nitrate concentration to trigger proper metabolic acclimation. CHL1 has therefore become a paradigm for how nutrient transporters may also serve as nutrient sensors, and how optimization of transport and signalling can be used simultaneously to improve nutrient efficiency.

Using dual-affinity binding, and with the help of two protein kinases (CIPK23 and CIPK8), CHL1 senses a wide range of nitrate concentration changes in the soil to alter its own transport properties[75,76] (Fig. 5b). CHL1 and NRT2.1 are also important for nitrate-regulated root development[77,78]. Vigorous root development is important for plants to compete

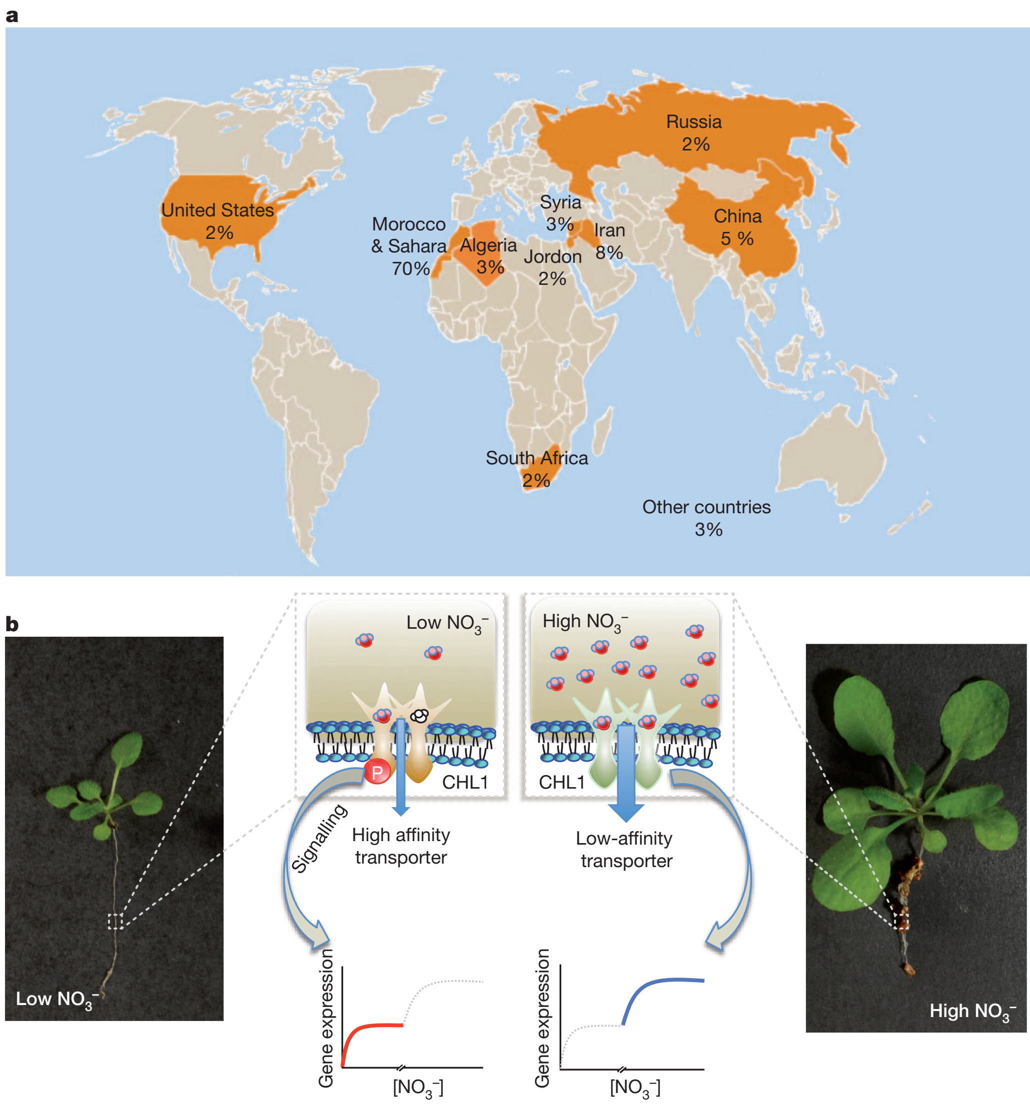


Figure 5 | Global phosphate availability and nitrate sensing. a, Global distribution map of reserves of rock phosphate. The percentage of effective reserves is illustrated (data from ref. 93). **b**, Dual functions of the nitrate transporter and nitrate sensor, CHL1, in nitrate uptake and sensing. Photographs show *Arabidopsis* seedlings at low (left) and high (right) nitrate. At low nitrate, CHL1 is phosphorylated at a specific amino acid, T101, converting it to a high-affinity nitrate uptake transporter, enabling nitrate accumulation at limiting soil nitrate concentrations. At high nitrate, the T101 amino acid is not phosphorylated, and CHL1 functions as a low-affinity transporter.

for nutrients and sustain crop yield[61,79]. Therefore, nitrate transporters and other proteins that regulate nitrate uptake and sensing provide potential tools for engineering crops with tailored N uptake activity, N metabolism and improved root growth for enhanced nitrogen-use efficiency and reduced-N-fertilizer requirements.

Future outlook

Our knowledge of the molecular nature and regulation of transporters has expanded vastly over the past twenty years. As shown in the examples here, fundamental research into transport mechanisms in plants is leading to rapid innovations for improving yields, extending the range of arable land that can be used for crops and improving the performance of plants under stress. This research also points to new solutions for more sustainable use of limited soil nutrients and for enhanced human nutrition through biofortification.

Recent advances on other plant transporters add to the list of potentially innovative applications in agriculture. For example, plants in the *Brassica* family, which includes oilseed rape (canola) and mustard, produce glucosinolates, which are potent defence compounds against herbivores and plant pathogens. A transporter controls the distribution of glucosinolates in *Arabidopsis*[80] and may be engineered to enhance herbivore resistance, as a way of reducing the application of pesticides. In another example, toxic heavy metal and arsenic accumulation in edible plant tissues, including rice in the United States, poses a threat to human health[81]. Plant transporters have recently been identified that control toxic metal and arsenic accumulation in seeds and other tissues[82–84], pointing directly to potential applications for developing plants with reduced levels of toxicant accumulation in grain and other edible tissues. Rice with nearly cadmium-free grain has been produced by identifying cadmium transporters and also using a molecular marker to select for genotypes that accumulate low cadmium concentrations[85].

Drought tolerance and desiccation avoidance by plants is critical for conserving water during drought periods. Plants lose over 90% of their water by transpiration from stomatal pores in the epidermis of leaves. Stomatal pores are also the gateways for CO_2 intake into plants for photosynthesis and have a key role in determining the water-use efficiency of plants. Research has shown that modification in the expression level of ion channels in guard cells (which control the opening and closing of stomatal pores) can be used to reduce water loss, enhance water-use efficiency and regulate efficient CO_2 intake for photosynthesis[86–88].

Moreover, transporters for the plant abiotic stress resistance hormone, abscisic acid, have been identified[89,90], and these may be used to target drought resistance responses. Targeting drought tolerance to particularly drought-sensitive tissues or organs during particularly susceptible stages of reproductive development (for example, grain filling and pollen meiosis) could become an important strategy during prolonged droughts and other predicted consequences of climate change.

Transport processes are key to photosynthesis. Since Peter Mitchell's groundbreaking 'chemiosmotic hypothesis' in 1961, the relevance of transporters in photosynthesis has become abundantly clear. Major advances have been made in identifying metabolite transporters across chloroplast membranes[91]. Yet many of the key transporters in chloroplast membranes remain unidentified. Discovery of the many predicted transporters in subcellular compartments, specifically in chloroplasts, has potential for improving energy capture.

A major challenge in future agriculture is establishing which genetic traits can be combined, or 'pyramided', without adversely affecting yield. Many transport processes reviewed here enhance plant performance via defined functions in specific tissues or cell types, so these may be particularly amenable to pyramiding. For instance, salinity tolerance that operates by removal of toxic sodium ions from the xylem sap[23–25,28] could be combined with traits that enhance sequestration of sodium into vacuoles[32,33], to confer additional salt tolerance. More work will be needed to determine whether or not traits will be compatible when combined. Moreover, many fundamental mechanisms for essential transport processes remain to be uncovered and many essential transporters undoubtedly remain to be discovered. Therefore, knowledge-targeted pyramiding of traits will require future advances in fundamental research into plant membrane transport processes.

Recent advances have identified plant membrane transporters and underlying mechanisms that increase the stress resistance and yield of staple crops. We expect that research into fundamental mechanisms of plant membrane transport processes will continue to produce surprises and breakthroughs that will provide new avenues towards a more sustainable and productive agriculture in the face of impending challenges.

Received 21 June 2012; accepted 11 January 2013.

1. The Food and Agriculture Organization of the United Nations. *The State of Food Insecurity in the World 2010* 8–11 (FAO, 2010).
2. The World Bank *Repositioning Nutrition as Central to Development: A Strategy for Large-Scale Action* Ch. 2, 42–61 (The International Bank for Reconstruction and Development/The World Bank, 2006); http://siteresources.worldbank.org/NUTRITION/Resources/281846-1131636806329/NutritionStrategy.pdf.
3. World Health Organization/FAO. *Diet, Nutrition and the Prevention of* Chronic Diseases 4–12 (2003); http://whqlibdoc.who.int/trs/who_trs_916.pdf.
4. *Foresight. The Future of Food and Farming: Final Project Report* 49–74 (The Government Office for Science, 2011), http://www.bis.gov.uk/assets/foresight/docs/food-and-farming/11-546-future-of-food-and-farming-report.pdf.
5. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461,** 472–475 (2009).
6. Foley, J. A. *et al.* Solutions for a cultivated planet. *Nature* **478,** 337–342 (2011).
7. Mueller, N. D. *et al.* Closing yield gaps through nutrient and water management. *Nature* **490,** 254–257 (2012).
8. Connor, D. J. Organic agriculture cannot feed the world. *Field Crops Res.* **106,** 187–190 (2008).
9. Conway, G. *One Billion Hungry: Can We Feed the World?* Ch. 7, 125–142 (Cornell Univ. Press, 2012).
10. Sanjana, N. E. *et al.* A transcription activator-like effector toolbox for genome engineering. *Nature Protocols* **7,** 171–192 (2012).
11. von Uexküll, H. R. & Mutert, E. Global extent, development and economic impact of acid soils. *Plant Soil* **171,** 1–15 (1995).
12. Ryan, P. R., Delhaize, E. & Jones, D. L. Function and mechanism of organic anion exudation from plant roots. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **52,** 527–560 (2001).
13. Sasaki, T. *et al.* A wheat gene encoding an aluminum-activated malate transporter. *Plant J.* **37,** 645–653 (2004).
These authors identified and characterized the plant aluminium tolerance protein, TaALMT1, an Al-activated anion channel that mediates the efflux of Al-detoxifying malate anion from the wheat root tip.
14. Delhaize, E. *et al.* Transgenic barley (*Hordeum vulgare L.*) expressing the wheat aluminium resistance gene (*TaALMT1*) shows enhanced phosphorus nutrition and grain production when grown on an acid soil. *Plant Biotechnol. J.* **7,** 391–400 (2009).
15. Rogers, E. E. & Guerinot, M. L. FRD3, a member of the multidrug and toxin efflux family, controls iron deficiency responses in *Arabidopsis. Plant Cell* **14,** 1787–1799 (2002).
16. Magalhaes, J. V. *et al.* A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. *Nature Genet.* **39,** 1156–1161 (2007).
17. Furukawa, J. *et al.* An aluminum-activated citrate transporter in barley. *Plant Cell Physiol.* **48,** 1081–1091 (2007).
18. Maron, L. G. *et al.* Two functionally distinct members of the MATE (multi-drug and toxic compound extrusion) family of transporters potentially underlie two major aluminum tolerance QTLs in maize. *Plant J.* **61,** 728–740 (2010).
19. Xia, J., Yamaji, N., Kasai, T. & Ma, J. F. Plasma membrane-localized transporter for aluminum in rice. *Proc. Natl Acad. Sci. USA* **107,** 18381–18385 (2010).
20. Huang, C.-F., Yamaji, N., Chen, Z. & Ma, J. F. A tonoplast-localized half-size ABC transporter is required for internal detoxification of aluminum in rice. *Plant J.* **69,** 857–867 (2012).
21. Famoso, A. N. *et al.* Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. *PLoS Genet.* **7,** e1002221 (2011).
22. Rubio, F., Gassmann, W. & Schroeder, J. I. Sodium-driven potassium uptake by the plant potassium transporter HKT1 and mutations conferring salt tolerance. *Science* **270,** 1660–1663 (1995).
23. Mäser, P. *et al.* Altered shoot/root Na^+ distribution and bifurcating salt sensitivity in *Arabidopsis* by genetic disruption of the Na^+ transporter AtHKT1. *FEBS Lett.* **531,** 157–161 (2002).
24. Ren, Z. H. *et al.* A rice quantitative trait locus for salt tolerance encodes a sodium transporter. *Nature Genet.* **37,** 1141–1146 (2005).
25. Sunarpi, *et al.* Enhanced salt tolerance mediated by AtHKT1 transporter-induced Na unloading from xylem vessels to xylem parenchyma cells. *Plant J.* **44,** 928–938 (2005).
26. Moller, I. S. *et al.* Shoot Na^+ exclusion and increased salinity tolerance engineered by cell type-specific alteration of Na^+ transport in *Arabidopsis. Plant Cell* **21,** 2163–2178 (2009).
27. Huang, S. *et al.* A sodium transporter (HKT7) is a candidate for *Nax1*, a gene for salt tolerance in durum wheat. *Plant Physiol.* **142,** 1718–1727 (2006).
28. Munns, R. *et al.* Wheat grain yield on saline soils is improved by an ancestral Na^+ transporter gene. *Nature Biotechnol.* **30,** 360–364 (2012).
A class 1 HKT transporter gene that prevents sodium accumulation in leaves was transferred from a wheat ancestor into modern durum wheat, with a resulting 25% increase in grain yield on saline soils.
29. Blumwald, E. & Poole, R. Na^+/H^+-antiport in isolated tonoplast vesicles from storage tissue of *Beta vulgaris. Plant Physiol.* **78,** 163–167 (1985).
30. Schachtman, D. P. & Schroeder, J. I. Structure and transport mechanism of a high-affinity potassium uptake transporter from higher plants. *Nature* **370,** 655–658 (1994).
31. Horie, T. *et al.* Rice OsHKT2;1 transporter mediates large Na^+ influx component into K^+-starved roots for growth. *EMBO J.* **26,** 3003–3014 (2007).
32. Apse, M. P., Aharon, G. S., Snedden, W. A. & Blumwald, E. Salt tolerance conferred by overexpression of a vacuolar Na^+/H^+ antiport in *Arabidopsis. Science* **285,** 1256–1258 (1999).
33. Barragan, V. *et al.* Ion exchangers NHX1 and NHX2 mediate active potassium uptake into vacuoles to regulate cell turgor and stomatal function in *Arabidopsis. Plant Cell* **24,** 1127–1142 (2012).
34. Mian, A. *et al.* Over-expression of an Na^+-and K^+-permeable HKT transporter in barley improves salt tolerance. *Plant J.* **68,** 468–479 (2011).
35. Riesmeier, J. W., Willmitzer, L. & Frommer, W. B. Evidence for an essential role of the sucrose transporter in phloem loading and assimilate partitioning. *EMBO J.* **13,** 1–7 (1994).
36. Chen, L. Q. *et al.* Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature* **468,** 527–532 (2010).
37. Chen, L. Q. *et al.* Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. *Science* **335,** 207–211 (2012).
A FRET (fluorescence (Förster) resonance energy transfer) sucrose nanosensor was used to identify the missing link in phloem loading, that is, the phloem-parenchyma-expressed SWEET sucrose transporters, which are also key susceptibility factors for bacterial pathogens in rice.
38. Patrick, J. W. Phloem unloading: sieve element unloading and post-sieve element transport. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48,** 191–222 (1997).
39. Antony, G. *et al.* Rice xa13 recessive resistance to bacterial blight is defeated by induction of the disease susceptibility gene Os-11N3. *Plant Cell* **22,** 3864–3876 (2010).
40. Chu, Z. *et al.* Targeting *xa13*, a recessive gene for bacterial blight resistance in rice. *Theor. Appl. Genet.* **112,** 455–461 (2006).
41. Chu, Z. *et al.* Promoter mutations of an essential gene for pollen development result in disease resistance in rice. *Genes Dev.* **20,** 1250–1255 (2006).
42. Liu, Q. *et al.* A paralog of the MtN3/saliva family recessively confers race-specific resistance to *Xanthomonas oryzae* in rice. *Plant Cell Environ.* **34,** 1958–1969 (2011).
43. Li, C., Wei, J., Lin, Y. & Chen, H. Gene silencing using the recessive rice bacterial blight resistance gene *xa13* as a new paradigm in plant breeding. *Plant Cell Rep.* **31,** 851–862 (2012).
44. Li, T., Liu, B., Spalding, M. H., Weeks, D. P. & Yang, B. High-efficiency TALEN-based gene editing produces disease-resistant rice. *Nature Biotechnol.* **30,** 390–392 (2012).
Synthetic transcriptional-activator-like effectors were used to develop rice plants that are resistant to the blight pathogen *Xanthomonas oryzae*, such that

the pathogen can no longer induce SWEET transporters, thus starving the pathogen by reducing the rice-derived sugar supply to the pathogen.

45. Masuda, H. *et al.* Increase in iron and zinc concentrations in rice grains via the introduction of barley genes involved in phytosiderophore synthesis. *Rice* **1,** 100–108 (2008).
46. Lee, S. *et al.* Iron fortification of rice seeds through activation of the nicotianamine synthase gene. *Proc. Natl Acad. Sci. USA* **106,** 22014–22019 (2009).
These authors showed that increasing nicotianamine levels resulted in increased levels of iron in polished rice and also that the iron is bioavailable using animal feeding studies.
47. Lee, S. *et al.* Activation of rice *nicotianamine synthase 2* (*OsNAS2*) enhances iron availability for biofortification. *Mol. Cells* **33,** 269–275 (2012).
48. Ishimaru, Y. *et al.* Rice metal-nicotianamine transporter, OsYSL2, is required for the long-distance transport of iron and manganese. *Plant J.* **62,** 379–390 (2010).
49. Wirth, J. *et al.* Rice endosperm iron biofortification by targeted and synergistic action of nicotianamine synthase and ferritin. *Plant Biotechnol. J.* **7,** 631–644 (2009).
50. Masuda, H. *et al.* Iron biofortification in rice by the introduction of multiple genes involved in iron nutrition. *Sci. Rep.* **2,** 543–549 (2012).
51. Lee, S. *et al.* Bio-available zinc in rice seeds is increased by activation tagging of nicotianamine synthase. *Plant Biotechnol. J.* **9,** 865–873 (2011).
52. Palmgren, M. G. *et al.* Zinc biofortification of cereals: problems and solutions. *Trends Plant Sci.* **13,** 464–473 (2008).
53. Lanquar, V. *et al.* Mobilization of vacuolar iron by AtNRAMP3 and AtNRAMP4 is essential for seed germination on low iron. *EMBO J.* **24,** 4041–4051 (2005).
54. Kim, S. A. *et al.* Localization of iron in *Arabidopsis* seed requires the vacuolar membrane transporter VIT1. *Science* **314,** 1295–1298 (2006).
These authors used X-ray fluorescence microprobe spectroscopy to localize iron in seeds and showed that failure to store iron properly in the vacuole via the VIT1 transporter leads to seedling lethality under iron limitation.
55. Morrissey, J. *et al.* The ferroportin metal efflux proteins function in iron and cobalt homeostasis in *Arabidopsis. Plant Cell* **21,** 3326–3338 (2009).
56. Zhang, Y., Xu, Y.-H., Yi, H.-Y. & Gong, J.-M. Vacuolar membrane transporters OsVIT1 and OsVIT2 modulate iron translocation between flag leaves and seeds in rice. *Plant J.* **72,** 400–410 (2012).
57. Podar, D. *et al.* Metal selectivity determinants in a family of transition metal transporters. *J. Biol. Chem.* **287,** 3185–3196 (2012).
58. Eide, D., Broderius, M., Fett, J. & Guerinot, M. L. A novel iron-regulated metal transporter from plants identified by functional expression in yeast. *Proc. Natl Acad. Sci. USA* **93,** 5624–5628 (1996).
59. Cakmak, I. Plant nutrition research: priorities to meet human needs for food in sustainable ways. *Plant Soil* **247,** 3–24 (2002).
60. López-Arredondo, D. L. & Herrera-Estrella, L. Engineering phosphorus metabolism in plants to produce a dual fertilization and weed control system. *Nature Biotechnol.* **30,** 889–893 (2012).
61. Gamuyao, R. *et al.* The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature* **488,** 535–539 (2012).
62. Beebe, S. E. *et al.* Quantitative trait loci for root architecture traits correlated with phosphorus acquisition in common bean. *Crop Sci.* **46,** 413–423 (2006).
63. Shin, R. & Schachtman, D. P. Hydrogen peroxide mediates plant root cell response to nutrient deprivation. *Proc. Natl Acad. Sci. USA* **101,** 8827–8832 (2004).
64. Remy, E. *et al.* The Pht1;9 and Pht1;8 transporters mediate inorganic phosphate acquisition by the *Arabidopsis thaliana* root during phosphorus starvation. *New Phytol.* **195,** 356–371 (2012).
65. Versaw, W. K. & Harrison, M. J. A chloroplast phosphate transporter, PHT2;1, influences allocation of phosphate within the plant and phosphate-starvation responses. *Plant Cell* **14,** 1751–1766 (2002).
66. Nagarajan, V. K. *et al. Arabidopsis* Pht1;5 mobilizes phosphate between source and sink organs and influences the interaction between phosphate homeostasis and ethylene signaling. *Plant Physiol.* **156,** 1149–1163 (2011).
67. Arpat, A. B. *et al.* Functional expression of PHO1 to the Golgi and trans-Golgi network and its role in export of inorganic phosphate. *Plant J.* **71,** 479–491 (2012).
68. Javot, H., Penmetsa, R. V., Terzaghi, N., Cook, D. R. & Harrison, M. J. A *Medicago truncatula* phosphate transporter indispensable for the arbuscular mycorrhizal symbiosis. *Proc. Natl Acad. Sci. USA* **104,** 1720–1725 (2007).
A phosphate transporter (MtPT4) was shown to be necessary for *Medicago truncatula* plants to obtain phosphate delivered via the fungal symbiont and furthermore, that MtPT4 transporter function is essential to maintain the symbiosis.
69. Yang, S. Y. *et al.* Nonredundant regulation of rice arbuscular mycorrhizal symbiosis by two members of the *PHOSPHATE TRANSPORTER1* gene family. *Plant Cell* **24,** 4236–4251 (2012).
70. McAllister, C. H., Beatty, P. H. & Good, A. G. Engineering nitrogen use efficient crop plants: the current status. *Plant Biotechnol. J.* **10,** 1011–1025 (2012).
71. Wang, Y. Y., Hsu, P. K. & Tsay, Y. F. Uptake, allocation and signaling of nitrate. *Trends Plant Sci.* **17,** 458–467 (2012).
72. Kiba, T. *et al.* The *Arabidopsis* nitrate transporter NRT2.4 plays a double role in roots and shoots of nitrogen-starved plants. *Plant Cell* **24,** 245–258 (2012).
73. Liu, K. H. & Tsay, Y. F. Switching between the two action modes of the dual-affinity nitrate transporter CHL1 by phosphorylation. *EMBO J.* **22,** 1005–1013 (2003).
74. Wang, R., Liu, D. & Crawford, N. M. The *Arabidopsis* CHL1 protein plays a major role in high-affinity nitrate uptake. *Proc. Natl Acad. Sci. USA* **95,** 15134–15139 (1998).
75. Ho, C. H., Lin, S. H., Hu, H. C. & Tsay, Y. F. CHL1 functions as a nitrate sensor in plants. *Cell* **138,** 1184–1194 (2009).
These authors provided the first report of a nutrient transporter in plants that also acts as a sensor for its own substrate, nitrate, over a wide range of concentrations.
76. Hu, H. C., Wang, Y. Y. & Tsay, Y. F. AtCIPK8, a CBL-interacting protein kinase, regulates the low-affinity phase of the primary nitrate response. *Plant J.* **57,** 264–278 (2009).
77. Little, D. Y. *et al.* The putative high-affinity nitrate transporter NRT2.1 represses lateral root initiation in response to nutritional cues. *Proc. Natl Acad. Sci. USA* **102,** 13693–13698 (2005).
78. Krouk, G. *et al.* Nitrate-regulated auxin transport by NRT1.1 defines a mechanism for nutrient sensing in plants. *Dev. Cell* **18,** 927–937 (2010).
79. Ruffel, S. *et al.* Nitrogen economics of root foraging: transitive closure of the nitrate-cytokinin relay and distinct systemic signaling for N supply vs. demand. *Proc. Natl Acad. Sci. USA* **108,** 18524–18529 (2011).
80. Nour-Eldin, H. H. *et al.* NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *Nature* **488,** 531–534 (2012).
81. Gilbert-Diamond, D. *et al.* Rice consumption contributes to arsenic exposure in US women. *Proc. Natl Acad. Sci. USA* **108,** 20656–20660 (2011).
82. Ueno, D. *et al.* Gene limiting cadmium accumulation in rice. *Proc. Natl Acad. Sci. USA* **107,** 16500–16505 (2010).
83. Song, W. Y. *et al.* Arsenic tolerance in *Arabidopsis* is mediated by two ABCC-type phytochelatin transporters. *Proc. Natl Acad. Sci. USA* **107,** 21187–21192 (2010).
84. Ma, J. F. *et al.* Transporters of arsenite in rice and their role in arsenic accumulation in rice grain. *Proc. Natl Acad. Sci. USA* **105,** 9931–9935 (2008).
These authors demonstrated that arsenite, a toxic form of soil arsenic, is transported into and within the rice plant by two novel transporters, providing new possible strategies for minimizing arsenic entry into the food chain by altering these transporters.
85. Ishikawa, S. *et al.* Ion-beam irradiation, gene identification, and marker-assisted breeding in the development of low-cadmium rice. *Proc. Natl Acad. Sci. USA* **109,** 19166–19171 (2012).
86. Kim, T.-H., Böhmer, M., Hu, H., Nishimura, N. & Schroeder, J. I. Guard cell signal transduction network: advances in understanding abscisic acid, CO_2, and Ca^{2+} signaling. *Annu. Rev. Plant Biol.* **61,** 561–591 (2010).
87. Kusumi, K., Hirotsuka, S., Kumamaru, T. & Iba, K. Increased leaf photosynthesis caused by elevated stomatal conductance in a rice mutant deficient in SLAC1, a guard cell anion channel protein. *J. Exp. Bot.* **63,** 5635–5644 (2012).
88. Hu, H. *et al.* Carbonic anhydrases are upstream regulators of CO_2-controlled stomatal movements in guard cells. *Nature Cell Biol.* **12,** 87–93 (2010).
89. Kuromori, T. *et al.* ABC transporter AtABCG25 is involved in abscisic acid transport and responses. *Proc. Natl Acad. Sci. USA* **107,** 2361–2366 (2010).
90. Kang, J. *et al.* PDR-type ABC transporter mediates cellular uptake of the phytohormone abscisic acid. *Proc. Natl Acad. Sci. USA* **107,** 2355–2360 (2010).
91. Weber, A. P. & Brautigam, A. The role of membrane transport in metabolic engineering of plant primary metabolism. *Curr. Opin. Biotechnol.* http://dx.doi.org/10.1016/j.copbio.2012.09.010 (4 October 2012).
92. Delhaize, E., Gruber, B. D. & Ryan, P. R. The roles of organic anion permeases in aluminium tolerance and mineral nutrition. *FEBS Lett.* **581,** 2255–2262 (2007).
93. US Geological Survey *Mineral Commodity Summaries 2012* 118–119 (US Geological Survey, 2012); http://minerals.usgs.gov/minerals/pubs/mcs/2012/mcs2012.pdf.

Acknowledgements Research in our laboratories was supported by: the Division of Chemical Sciences, Geosciences and Biosciences, Office of Basic Energy Sciences at the US Department of Energy (DOE) under grant numbers DE-FG02-03ER15449 (to J.I.S.), DE-FG02-04ER15542 (to W.B.F.) and DE-FG-2-06ER15809 (to M.L.G.); by the Grains Research and Development Corporation, Australia (to R.M. and E.D.); by the US National Science Foundation under grant numbers IOS:0842720 (to M.J.H.), MCB0918220 (to J.I.S.) and IOS-091994 and DBI 0701119 (to M.L.G.); by the UK Biotechnology and Biological Sciences Research Council under grant number BB/J004561/1 (to D.S.); by the National Institutes of Health under grant numbers GM060396-P42ES010337 (to J.I.S.) and GM078536 and P42ES007373 (to M.L.G.); by the US Department of Agriculture under grant number 2009-02273 (to L.V.K.); by a Generation Challenge Grant under grant number G7010.03.06 (to L.V.K.); by the Howard Hughes Medical Institute under grant number 55005946 (to L.H.E.); the CREST Japan Science and Technology Agency (to N.K.N.); by the Ministry of Education, Culture, Sports, Science and Technology, Japan under grant number 23119507 (to T.H.); and by the Academia Sinica, Taiwan and the National Science Council, Taiwan under grant number NSC 101-2321-B-001-005 (to Y.F.T.).

Author Contributions The project was conceived and outlined by J.I.S. The manuscript was planned by J.I.S. and D.S. All authors contributed to writing sections of the manuscript and all authors commented on versions of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.I.S. (jischroeder@ucsd.edu) or D.S. (dale.sanders@jic.ac.uk).

OPEN
doi:10.1038/nature12113

Integrated genomic characterization of endometrial carcinoma

The Cancer Genome Atlas Research Network*

We performed an integrated genomic, transcriptomic and proteomic characterization of 373 endometrial carcinomas using array- and sequencing-based technologies. Uterine serous tumours and ~25% of high-grade endometrioid tumours had extensive copy number alterations, few DNA methylation changes, low oestrogen receptor/progesterone receptor levels, and frequent *TP53* mutations. Most endometrioid tumours had few copy number alterations or *TP53* mutations, but frequent mutations in *PTEN*, *CTNNB1*, *PIK3CA*, *ARID1A* and *KRAS* and novel mutations in the SWI/SNF chromatin remodelling complex gene *ARID5B*. A subset of endometrioid tumours that we identified had a markedly increased transversion mutation frequency and newly identified hotspot mutations in *POLE*. Our results classified endometrial cancers into four categories: *POLE* ultramutated, microsatellite instability hypermutated, copy-number low, and copy-number high. Uterine serous carcinomas share genomic features with ovarian serous and basal-like breast carcinomas. We demonstrated that the genomic features of endometrial carcinomas permit a reclassification that may affect post-surgical adjuvant treatment for women with aggressive tumours.

Endometrial cancer arises from the lining of the uterus. It is the fourth most common malignancy among women in the United States, with an estimated 49,500 new cases and 8,200 deaths in 2013 (ref. 1). Most patients present with low-grade, early-stage disease. The majority of patients with more aggressive, high-grade tumours who have disease spread beyond the uterus will progress within 1 year (refs 2, 3). Endometrial cancers have been broadly classified into two groups[4]. Type I endometrioid tumours are linked to oestrogen excess, obesity, hormone-receptor positivity, and favourable prognosis compared with type II, primarily serous, tumours that are more common in older, non-obese women and have a worse outcome. Early-stage endometrioid cancers are often treated with adjuvant radiotherapy, whereas serous tumours are treated with chemotherapy, similar to advanced-stage cancers of either histological subtype. Therefore, proper subtype classification is crucial for selecting appropriate adjuvant therapy.

Several previous reports suggest that *PTEN* mutations occur early in the neoplastic process of type I tumours and co-exist frequently with other mutations in the phosphatidylinositol-3-OH kinase (PI(3)K)/AKT pathway[5,6]. Other commonly mutated genes in type I tumours include *FGFR2*, *ARID1A*, *CTNNB1*, *PIK3CA*, *PIK3R1* and *KRAS*[7–9]. Microsatellite instability (MSI) is found in approximately one-third of type I tumours, but is infrequent in type II tumours[10]. *TP53*, *PIK3CA* and *PPP2R1A* mutations are frequent in type II tumours[11,12]. Most of these studies have been limited to DNA sequencing only with samples of heterogeneous histological subtypes and tumour grades. We present a comprehensive, multiplatform analysis of 373 endometrial carcinomas including low-grade endometrioid, high-grade endometrioid, and serous carcinomas. This integrated analysis provides key molecular insights into tumour classification, which may have a direct effect on treatment recommendations for patients, and provides opportunities for genome-guided clinical trials and drug development.

Results

Tumour samples and corresponding germline DNA were collected from 373 patients, including 307 endometrioid and 66 serous (53) or mixed histology (13) cases. Local Institutional Review Boards approved all tissue acquisition. The clinical and pathological characteristics of the samples generally reflect a cross-section of individuals with recurrent endometrial cancer[2,3] (Supplementary Table 1.1). The median follow-up of the cohort was 32 months (range, 1–19 months); 21% of the patients have recurred, and 11% have died. Comprehensive molecular analyses were performed at independent centres using six genomic or proteomic platforms (Supplementary Table 1.2). MSI testing performed on all samples using seven repeat loci (Supplementary Table 1.3) found MSI in 40% of endometrioid tumours and 2% of serous tumours.

Somatic copy number alterations

Somatic copy number alterations (SCNAs) were assessed in 363 endometrial carcinomas. Unsupervised hierarchical clustering grouped the tumours into four clusters (Fig. 1a). The first three copy-number clusters were composed almost exclusively (97%) of endometrioid tumours without significant differences in tumour grades. Cluster 1 tumours were nearly devoid of broad SCNAs, averaging less than 0.5% genome alteration, with no significant recurrent events. Cluster 1 tumours also had significantly increased non-synonymous mutation rates compared to all others (median 7.2×10^{-6} versus 1.7×10^{-6} mutations per megabase (Mb), $P < 0.001$). Copy-number clusters 2 and 3 consisted mainly of endometrioid tumours, distinguished by more frequent 1q amplification in cluster 3 than cluster 2 (100% of cluster 3 tumours versus 33% of cluster 2 tumours) and worse progression-free survival ($P = 0.003$, log-rank versus clusters 1 and 2; Fig. 1b).

Most of the serous (50 out of 53; 94%) and mixed histology (8 out of 13; 62%) tumours clustered with 36 (12%) of the 289 endometrioid tumours, including 24% of grade 3 and 5% of grade 1 or 2, into copy-number cluster 4; a single group characterized by a very high degree of SCNAs (Supplementary Fig. 2.1; focal SCNAs with false discovery rate (FDR) < 0.15, and Supplementary Data 2.1). Cluster 4 tumours were characterized by significantly recurrent previously reported focal amplifications of the oncogenes *MYC* (8q24.12), *ERBB2* (17q12) and *CCNE1* (19q12)[13], and by SCNAs previously unreported in endometrial cancers including those containing *FGFR3* (4p16.3) and *SOX17* (8q11.23). Cluster 4 tumours also had frequent *TP53* mutations (90%),

*Lists of participants and their affiliations appear at the end of the paper.

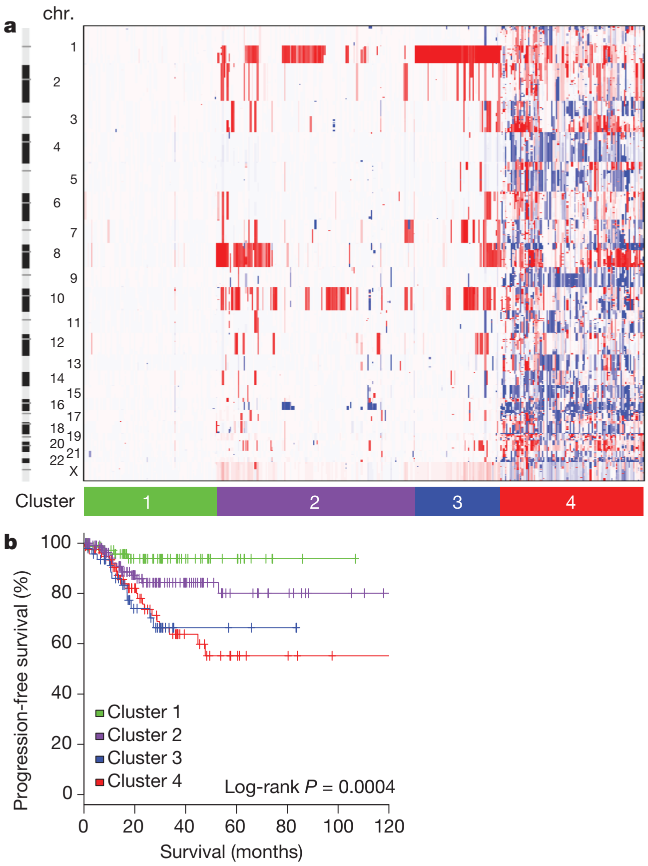


Figure 1 | SCNAs in endometrial carcinomas. **a**, Tumours were hierarchically clustered into four groups based on SCNAs. The heat map shows SCNAs in each tumour (horizontal axis) plotted by chromosomal location (vertical axis). Chr., chromosome. **b**, Kaplan–Meier curves of progression-free survival for each copy-number cluster.

little MSI (6%), and fewer *PTEN* mutations (11%) than other endometrioid tumours (84%). Overall, these findings suggest that a subset of endometrial tumours contain distinct patterns of SCNAs and mutations that do not correlate with traditional tumour histology or grade.

As expected, tumours in the 'serous-like' cluster (cluster 4) had significantly worse progression-free survival than tumours in the endometrioid cluster groups ($P = 0.003$, log-rank, Fig. 1b). Potential therapeutically relevant SCNAs included the cluster 2 15q26.2 focal amplification, which contained *IGF1R*; and cluster 4 amplifications of *ERBB2*, *FGFR1* and *FGFR3*, and *LRP1B* deletion, which was recently associated with resistance to liposomal doxorubicin in serous ovarian cancer[14].

Exome sequence analysis

We sequenced the exomes of 248 tumour/normal pairs. On the basis of a combination of somatic nucleotide substitutions, MSI and SCNAs, the endometrial tumours were classified into four groups (Fig. 2a, b): (1) an ultramutated group with unusually high mutation rates (232×10^{-6} mutations per Mb) and a unique nucleotide change spectrum; (2) a hypermutated group (18×10^{-6} mutations per Mb) of MSI tumours, most with *MLH1* promoter methylation; (3) a group with lower mutation frequency (2.9×10^{-6} mutations per Mb) and most of the microsatellite stable (MSS) endometrioid cancers; and (4) a group that consists primarily of serous-like cancers with extensive SCNA (copy-number cluster 4) and a low mutation rate (2.3×10^{-6} mutations per Mb). The ultramutated group consisted of 17 (7%) tumours exemplified by an increased C→A transversion frequency, all with mutations in the exonuclease domain of *POLE*, and an improved progression-free survival (Fig. 2a, c). *POLE* is a catalytic subunit of DNA polymerase epsilon involved in nuclear DNA replication and repair. We identified hotspot mutations in *POLE* at Pro286Arg and Val411Leu present in 13 (76%) of the 17 ultramutated samples. Significantly mutated genes (SMGs) identified at low FDRs (Q) in this subset included *PTEN* (94%, $Q = 0$), *PIK3R1* (65%, $Q = 8.3 \times 10^{-7}$), *PIK3CA* (71%, $Q = 9.1 \times 10^{-5}$), *FBXW7* (82%, $Q = 1.4 \times 10^{-4}$), *KRAS* (53%, $Q = 9.2 \times 10^{-4}$) and *POLE* (100%, $Q = 4.2 \times 10^{-3}$). Mutation rates in *POLE* mutant endometrial and previously reported ultramutated colorectal tumours exceeded those found in any other lineage including lung cancer and melanoma[15–17]. Germline susceptibility variants have been reported in *POLE* (Leu424Val) and *POLD1* (Ser478Asn), but were not found in our endometrial normal exome-seq reads[18].

The MSI endometrioid tumours had a mutation frequency approximately tenfold greater than MSS endometrioid tumours, few SCNAs, frameshift deletions in *RPL22*, frequent non-synonymous *KRAS* mutations, and few mutations in *FBXW7*, *CTNNB1*, *PPP2R1A* and *TP53*. The MSS, copy-number low, endometrioid tumours had an unusually high frequency of *CTNNB1* mutations (52%); the only gene with a higher mutation frequency than the MSI samples. The copy-number high group contained all of the remaining serous cases and one-quarter of the grade 3 endometrioid cases. Most of these tumours had *TP53* mutations and a high frequency of *FBXW7* (22%, $Q = 0$) and *PPP2R1A* (22%, $Q = 1.7 \times 10^{-16}$) mutations, previously reported as common in uterine serous but not endometrioid carcinomas. Thus, a subset of high-grade endometrioid tumours had similar SCNAs and mutation spectra as uterine serous carcinomas, suggesting that these patients might benefit from treatment approaches that parallel those for serous tumours.

There were 48 genes with differential mutation frequencies across the four groups (Fig. 2d and Supplementary Data 3.1). *ARID5B*, a member of the same AT-rich interaction domain (ARID) family as *ARID1A*, was more frequently mutated in MSI (23.1%) than in either MSS endometrioid (5.6%) or high SCNA serous tumours (0%), a novel finding for endometrial cancer. Frameshifting *RPL22* indels near a homopolymer at Lys 15 were almost exclusively found in the MSI group (36.9%). The *TP53* mutation frequency (>90%) in serous tumours differentiated them from the endometrioid subtypes (11.4%). However, many (10 out of 20; 50%) endometrioid tumours with a non-silent *TP53* mutation also had non-silent mutations in *PTEN*, compared to only 1 out of 39 (2.6%) serous tumours with non-silent *TP53* mutations. Although *TP53* mutations are not restricted to serous tumours, the co-existing *PTEN* mutations in the endometrioid cases suggest a distinct tumorigenic mechanism.

Comparisons of 66 SMGs between traditional histological subtypes are provided (Supplementary Methods 3), and SMGs across other subcohorts can be found in Supplementary Data 3.2. The spectrum of *PIK3CA* and *PTEN* mutations in endometrial cancer also differed from other solid tumours (Supplementary Methods 3). Integrated analysis may be useful for identifying histologically misclassified cases. For example, a single serous case was identified without a *TP53* mutation or extensive SCNAs and with a *KRAS* mutation and high mutation rate. After re-review of the histological section, the case was deemed consistent with a grade 3 endometrioid tumour, demonstrating how molecular analysis could reclassify tumour histology and potentially affect treatment decisions.

Multiplatform subtype classifications

All of the endometrial tumours were examined for messenger RNA expression ($n = 333$), protein expression ($n = 293$), microRNA expression ($n = 367$), and DNA methylation ($n = 373$) (Supplementary Methods 4–7). Unsupervised k-means clustering of mRNA expression from RNA sequencing identified three robust clusters termed 'mitotic', 'hormonal' and 'immunoreactive' (Supplementary Fig. 4.1) that were significantly correlated with the four integrated clusters; *POLE*, MSI, copy-number low and copy-number high ($P < 0.0001$). Supervised analysis identified signature genes of the *POLE* cluster ($n = 17$) mostly involved in cellular metabolism (Fig. 3a). Among the few signature genes

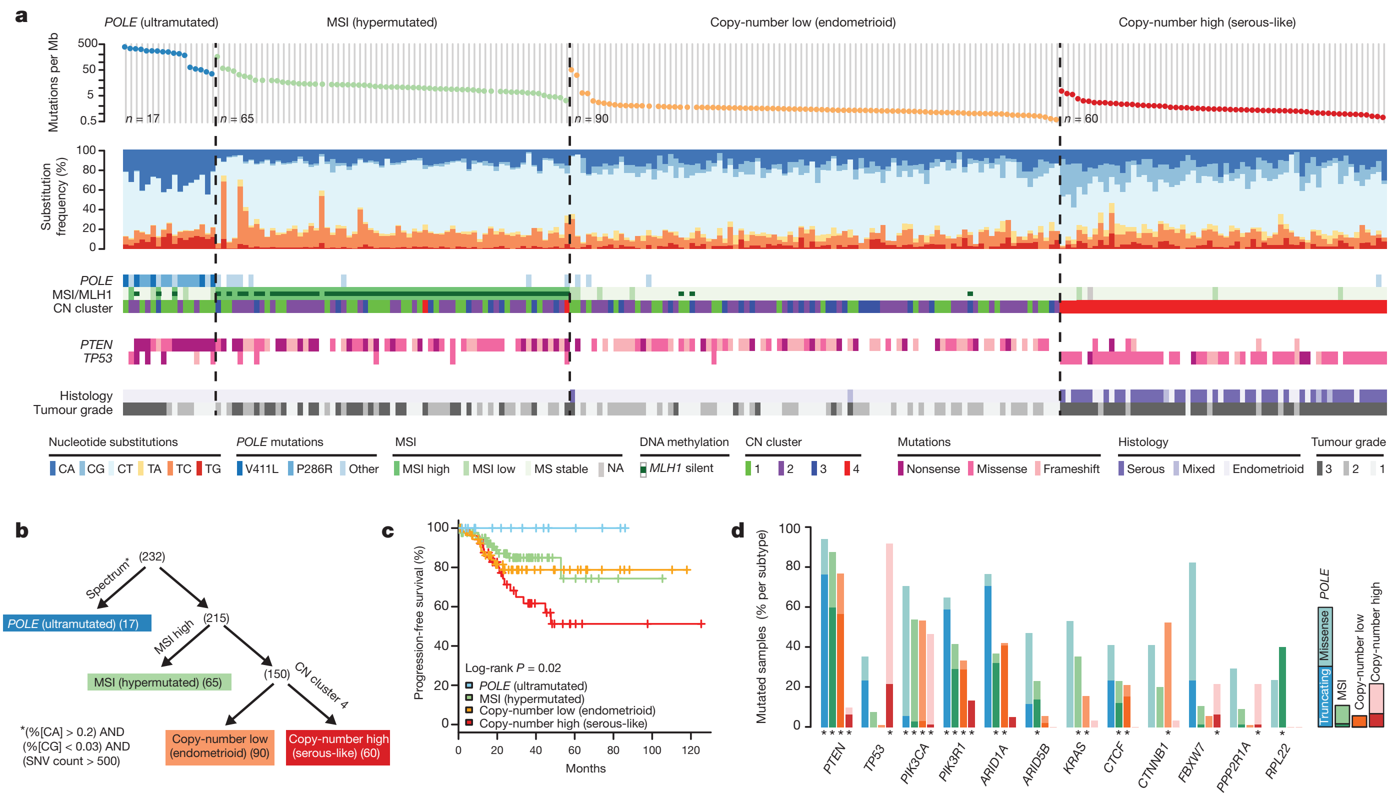


Figure 2 | Mutation spectra across endometrial carcinomas. **a**, Mutation frequencies (vertical axis, top panel) plotted for each tumour (horizontal axis). Nucleotide substitutions are shown in the middle panel, with a high frequency of C-to-A transversions in the samples with *POLE* exonuclease mutations. CN, copy number. **b**, Tumours were stratified into the four groups by (1) nucleotide substitution frequencies and patterns, (2) MSI status, and (3) copy-number cluster. SNV, single nucleotide variant. **c**, *POLE*-mutant tumours have significantly better progression-free survival, whereas copy-number high tumours have the poorest outcome. **d**, Recurrently mutated genes are different between the four subgroups. Shown are the mutation frequencies of all genes that were significantly mutated in at least one of the four subgroups (MUSiC, asterisk denotes FDR < 0.05).

in the MSI cluster was decreased *MLH1* mRNA expression, probably due to its promoter methylation. Increased progesterone receptor (*PGR*) expression was noted in the copy-number low cluster, suggesting responsiveness to hormonal therapy. The copy-number high cluster, which included most of the serous and serous-like endometrioid tumours, exhibited the greatest transcriptional activity exemplified by increased cell cycle deregulation (for example, *CCNE1*, *PIK3CA*, *MYC* and *CDKN2A*) and *TP53* mutation (Supplementary Figs 4.2 and 4.3). This is consistent with reports that increased CDKN2A can distinguish serous from endometrioid carcinomas[19]. Approximately 85% of cases in the copy-number high cluster shared membership with the 'mitotic' mRNA subtype.

Supervised clustering of the reverse phase protein array (RPPA) expression data was consistent with loss of function for many of the mutated genes (Fig. 3b). *TP53* was frequently mutated in the copy-number high group ($P = 2.5 \times 10^{-27}$) and its protein expression was also increased, suggesting that these mutations are associated with increased expression. By contrast, *PTEN* ($P = 2.8 \times 10^{-19}$) and *ARID1A* ($P = 1.2 \times 10^{-6}$) had high mutation rates in the remaining groups, but their expression was decreased, suggesting inactivating mutations in both genes. The copy-number high group also had decreased levels of phospho-AKT, consistent with downregulation of the AKT pathway. The copy-number low group had raised RAD50 expression, which is associated with DNA repair, explaining some of the differences between the copy-number high and low groups. The *POLE* group had high expression of ASNS and CCNB1, whereas the MSI tumours had both high phospho-AKT and low PTEN expression.

Unsupervised clustering of DNA methylation data generated from Illumina Infinium DNA methylation arrays revealed four unique subtypes (MC1–4) that support the four integrative clusters. A heavily methylated subtype (MC1) reminiscent of the CpG island methylator phenotype (CIMP) described in colon cancers and glioblastomas[20–22] was associated with the MSI subtype and attributable to promoter hypermethylation of *MLH1*. A serous-like cluster (MC3) with minimal DNA methylation changes was composed primarily of serous tumours and some endometrioid tumours (Supplementary Fig. 7.1) and contained most of the copy-number high tumours.

Integrative clustering using the iCluster framework returned two major clusters split primarily on serous and endometrioid histology highlighting *TP53* mutations, lack of *PTEN* mutation and encompassing almost exclusively copy-number high tumours[23] (Supplementary Fig. 8.1). We developed a new clustering algorithm, called SuperCluster, to derive overall subtypes based on sample cluster memberships across all data types (Supplementary Fig. 9.1). SuperCluster identified four clusters that generally confirmed the contributions of individual platforms to the overall integrated clusters. No major batch effects were identified for any platform (Supplementary Methods 10).

Structural aberrations

To identify somatic chromosomal aberrations, we performed low-pass, paired-end, whole-genome sequencing on 106 tumours with matched normals. We found recurrent translocations involving genes in several pathways including WNT, EGFR–RAS–MAPK, PI(3)K, protein kinase A, retinoblastoma and apoptosis. The most frequent translocations (5 out of 106) involved a member of the BCL family (*BCL2*, *BCL7A*, *BCL9* and *BCL2L11*). Four of these were confirmed by identification of the translocation junction point and two were also confirmed by high-throughput RNA sequencing (RNA-Seq). In all cases the translocations result in in-frame fusions and are predicted to result in activation or increased expression of the BCL family members (Supplementary Fig. 3.2). Translocations involving members of the BCL family leading to reduced apoptosis have been

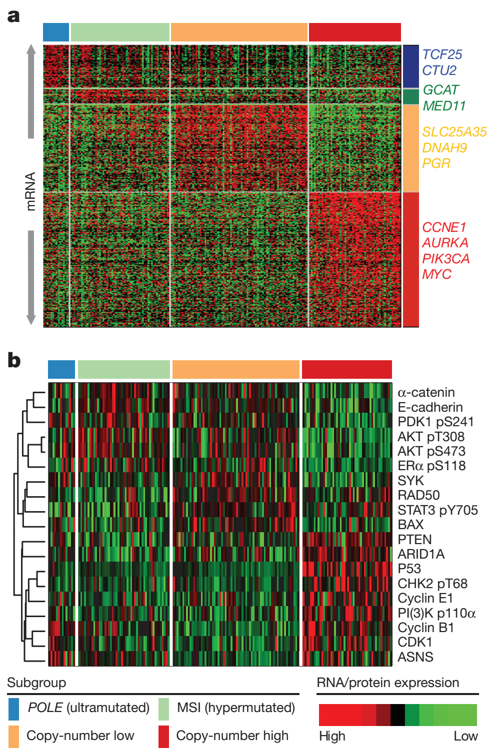


Figure 3 | Gene expression across integrated subtypes in endometrial carcinomas. **a**, Supervised analysis of ~1,500 genes significantly associated with integrated subtypes. **b**, Heat map of protein expression clusters, supervised by integrated subtypes. Samples are in columns; genes or proteins are in rows.

described in other tumour types[24] and our results suggest that similar mechanisms may be operative here.

Pathway alterations

Multiple platform data were integrated to identify recurrently altered pathways in the four endometrial cancer integrated subgroups. Because of the high background mutation rate and small sample size, we excluded the *POLE* subgroup from this analysis. Considering all recurrently mutated, homozygously deleted, and amplified genes, we used MEMo[25] to identify gene networks with mutually exclusive alteration patterns in each subgroup. The most significant module was found in the copy-number low group and contained *CTNNB1*, *KRAS* and *SOX17* (Fig. 4a). The very strong mutual exclusivity between mutations in these three genes suggests that alternative mechanisms activate WNT signalling in endometrioid endometrial cancer. Activating *KRAS* mutations have been shown to increase the stability of β-catenin via glycogen synthase kinase 3β (GSK-3β), leading to an alternative mechanism of β-catenin activation other than adenomatous polyposis coli degradation[26]. *SOX17*, which mediates proteasomal degradation of β-catenin[27,28], is mutated exclusively in the copy-number low group (8%) at recurrent positions (Ala96Gly and Ser403Ile) not previously described. Other genes with mutually exclusive alteration patterns in this module were *FBXW7*, *FGFR2* and *ERBB2* (ref. 29). *ERBB2* was focally amplified with protein overexpression in 25% of the serous or serous-like tumours, suggesting a potential role for human epidermal growth factor receptor 2 (HER2)-targeted inhibitors. A small clinical trial of trastuzumab found no activity in endometrial carcinoma, but accrued few HER2 fluorescence *in situ* hybridization (FISH)-amplified serous carcinomas[30].

PIK3CA and *PIK3R1* mutations were frequent and showed a strong tendency for mutual exclusivity in all subgroups, but unlike other tumour types, they co-occurred with *PTEN* mutations in the MSI and copy-number low subgroups as previously reported[5,9] (Fig. 4b). The copy-number high subgroup showed mutual exclusivity between alterations of all three genes. Overall, 93% of endometrioid tumours had mutations that suggested potential for targeted therapy with PI(3)K/AKT pathway inhibitors.

Consensus clustering of copy number, mRNA expression and pathway interaction data for 324 samples yielded five PARADIGM clusters with distinct pathway activation patterns[31] (Fig. 4c and Supplementary Methods 11). PARADIGM cluster 1 had the lowest level of MYC pathway activation and highest level of WNT pathway activation, consistent with its composition of copy-number low cases having frequent *CTNNB1* mutations. PARADIGM cluster 3 was composed predominantly of the copy-number high cases, with relatively high MYC/MAX signalling but low oestrogen receptor/FOXA1 signalling and p53 activity. Only *TP53* truncation and not missense mutations were implicated as loss-of-function mutations, suggesting different classes of p53 mutations may have distinct signalling consequences. PARADIGM cluster 5 was enriched for hormone receptor expression.

Comparison to ovarian and breast cancers

The clinical and pathologic features of uterine serous carcinoma and high-grade serous ovarian carcinoma (HGSOC) are quite similar. HGSOC shares many similar molecular features with basal-like breast carcinoma[32]. Focal SCNA patterns were similar between these three tumour subtypes and unsupervised clustering identified relatedness (Fig. 5a and Supplementary Fig. 12.1). Supervised analysis of transcriptome data sets showed high correlation between tumour subtypes (Supplementary Fig. 12.2). The MC3 DNA methylation subtype with minimal DNA methylation changes was also similar to basal-like breast and HGSOCs (Supplementary Fig. 12.3). A high frequency of *TP53* mutations is shared across these tumour subtypes (uterine serous, 91%; HGSOC, 96%; basal-like breast, 84%)[33,34], as is the very low frequency of *PTEN* mutations (uterine serous, 2%; HGSOC, 1%; basal-like breast, 1%). Differences included a higher frequency of *FBXW7*, *PPP2R1A* and *PIK3CA* mutations in uterine serous compared to basal-like breast and HGSOCs (Fig. 5b). We showed that uterine serous carcinomas share many molecular features with both HGSOCs and basal-like breast carcinomas, despite more frequent mutations, suggesting new opportunities for overlapping treatment paradigms.

Discussion

This integrated genomic and proteomic analysis of 373 endometrial cancers provides insights into disease biology and diagnostic classification that could have immediate therapeutic application. Our analysis identified four new groups of tumours based on integrated genomic data, including a novel *POLE* subtype in ~10% of endometrioid tumours. Ultrahigh somatic mutation frequency, MSS, and common, newly identified hotspot mutations in the exonuclease domain of *POLE* characterize this subtype. SCNAs add a layer of resolution, revealing that most endometrioid tumours have few SCNAs, most serous and serous-like tumours exhibit extensive SCNAs, and the extent of SCNA roughly correlates with progression-free survival.

Endometrial cancer has more frequent mutations in the PI(3)K/AKT pathway than any other tumour type studied by The Cancer Genome Atlas (TCGA) so far. Endometrioid endometrial carcinomas share many characteristics with colorectal carcinoma including a high frequency of MSI (40% and 11%, respectively), *POLE* mutations (7% and 3%, respectively) leading to ultrahigh mutation rates, and frequent activation of WNT/CTNNB1 signalling; yet endometrial carcinomas have novel exclusivity of *KRAS* and *CTNNB1* mutations and a distinct mechanism of pathway activation. Uterine serous carcinomas share many similar characteristics with basal-like breast and HGSOCs; three tumour types with high-frequency non-silent *TP53* mutations and extensive SCNA. However, the high frequency of *PIK3CA*, *FBXW7*, *PPP2R1A* and *ARID1A* mutations in uterine serous carcinomas are not found in basal-like breast and HGSOCs. The frequency of mutations in *PIK3CA*, *FBXW7* and *PPP2R1A* was ~30% higher than in a recently

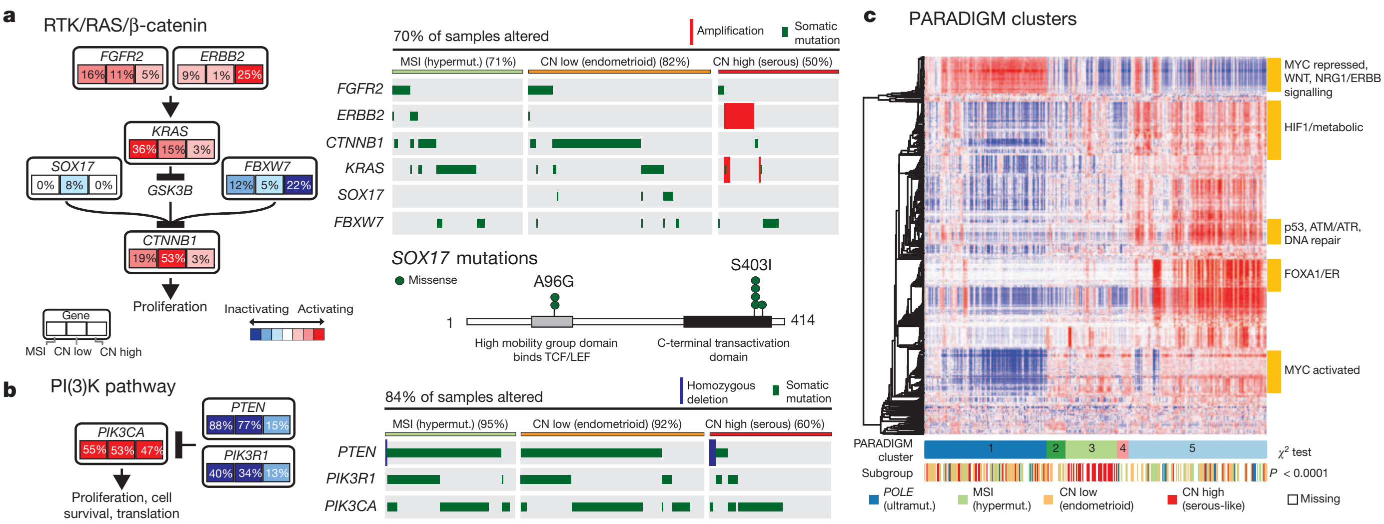


Figure 4 | Pathway alterations in endometrial carcinomas. **a**, The RTK/RAS/β-catenin pathway is altered through several mechanisms that exhibit mutually exclusive patterns. Alteration frequencies are expressed as a percentage of all cases. The right panel shows patterns of occurrence. **b**, The PI(3)K pathway has mutually exclusive *PIK3CA* and *PIK3R1* alterations that frequently co-occur with *PTEN* alterations in the MSI and copy-number low subgroups. **c**, Heat map display of top 1,000 varying pathway features within PARADIGM consensus clusters. Samples were arranged in order of their consensus cluster membership. The genomic subtype for each sample is displayed below the consensus clusters.

reported study of 76 uterine serous carcinomas[11], but similar to another study[12]. Uterine serous carcinomas have *ERBB2* amplification in 27% of tumours and *PIK3CA* mutations in 42%, which provide translational opportunities for targeted therapeutics.

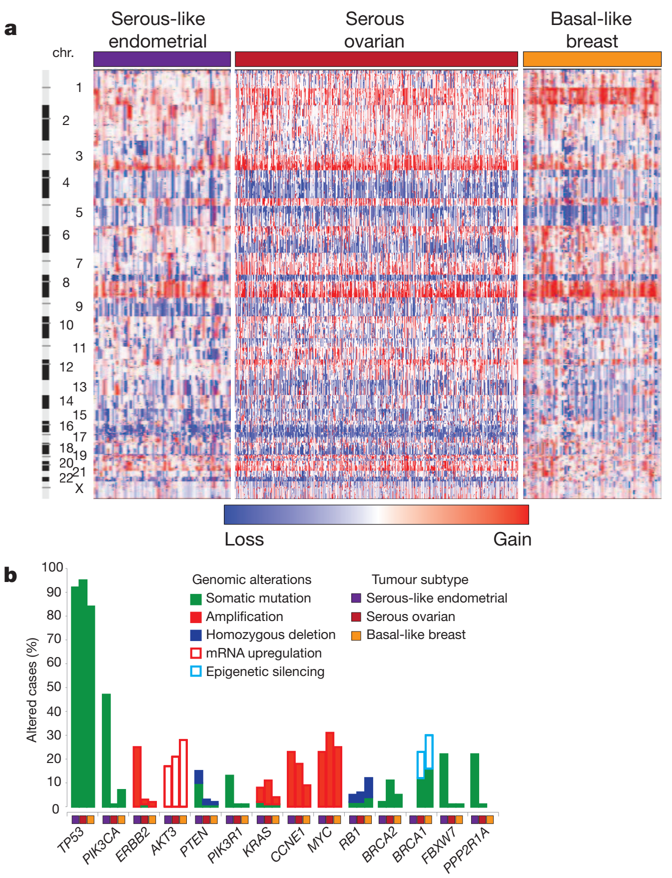


Figure 5 | Genomic relationships between endometrial serous-like, ovarian serous, and basal-like breast carcinomas. **a**, SCNAs for each tumour type. **b**, Frequency of genomic alterations present in at least 10% of one tumour type.

Early stage type I endometrioid tumours are often treated with adjuvant radiotherapy, whereas similarly staged type II serous tumours are treated with chemotherapy. High-grade serous and endometrioid endometrial carcinomas are difficult to subtype correctly, and intra-observer concordance among speciality pathologists is low[7,34–36]. Our molecular characterization data demonstrate that ~25% of tumours classified as high-grade endometrioid by pathologists have a molecular phenotype similar to uterine serous carcinomas, including frequent *TP53* mutations and extensive SCNA. The compelling similarities between this subset of endometrioid tumours and uterine serous carcinomas suggest that genomic-based classification may lead to improved management of these patients. Clinicians should carefully consider treating copy-number-altered endometrioid patients with chemotherapy rather than adjuvant radiotherapy and formally test such hypotheses in prospective clinical trials. Furthermore, the marked molecular differences between endometrioid and serous-like tumours suggest that these tumours warrant separate clinical trials to develop the independent treatment paradigms that have improved outcomes in other tumour types, such as breast cancer.

METHODS SUMMARY

Biospecimens were obtained from 373 patients after Institutional Review Board-approved consents. DNA and RNA were co-isolated using a modified AllPrep kit (Qiagen). We used Affymetrix SNP 6.0 microarrays to detect SCNAs in 363 samples and GISTIC analysis to identify recurrent events[37]. The exomes of 248 tumours were sequenced to a read-depth of at least ×20. We performed low-pass whole-genome sequencing on 107 tumours to a mean depth of ×6. Consensus clustering was used to analyse mRNA, miRNA, RPPA and methylation data with methods previously described[38–40]. Integrated cross-platform analyses were performed using MEMo, iCluster and PARADIGM[25,31].

Received 10 December 2012; accepted 21 March 2013.

1. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. *CA Cancer J. Clin.* **63,** 11–30 (2013).
2. Fleming, G. F. *et al.* Phase III trial of doxorubicin plus cisplatin with or without paclitaxel plus filgrastim in advanced endometrial carcinoma: a Gynecologic Oncology Group Study. *J. Clin. Oncol.* **22,** 2159–2166 (2004).
3. Sutton, G. *et al.* Whole abdominal radiotherapy in the adjuvant treatment of patients with stage III and IV endometrial cancer: a gynecologic oncology group study. *Gynecol. Oncol.* **97,** 755–763 (2005).
4. Lax, S. F. & Kurman, R. J. A dualistic model for endometrial carcinogenesis based on immunohistochemical and molecular genetic analyses. *Verh. Dtsch. Ges. Pathol.* **81,** 228–232 (1997).

5. Cheung, L. W. *et al.* High frequency of *PIK3R1* and *PIK3R2* mutations in endometrial cancer elucidates a novel mechanism for regulation of PTEN protein stability. *Cancer Discov.* **1,** 170–185 (2011).
6. Levine, R. L. *et al. PTEN* mutations and microsatellite instability in complex atypical hyperplasia, a precursor lesion to uterine endometrioid carcinoma. *Cancer Res.* **58,** 3254–3258 (1998).
7. McConechy, M. K. *et al.* Use of mutation profiles to refine the classification of endometrial carcinomas. *J. Pathol.* **228,** 20–30 (2012).
8. Byron, S. A. *et al. FGFR2* point mutations in 466 endometrioid endometrial tumors: relationship with MSI, *KRAS*, *PIK3CA*, *CTNNB1* mutations and clinicopathological features. *PLoS ONE* **7,** e30801 (2012).
9. Urick, M. E. *et al. PIK3R1* (p85α) is somatically mutated at high frequency in primary endometrial cancer. *Cancer Res.* **71,** 4061–4067 (2011).
10. Zighelboim, I. *et al.* Microsatellite instability and epigenetic inactivation of *MLH1* and outcome of patients with endometrial carcinomas of the endometrioid type. *J. Clin. Oncol.* **25,** 2042–2048 (2007).
11. Kuhn, E. *et al.* Identification of molecular pathway aberrations in uterine serous carcinoma by genome-wide analyses. *J. Natl. Cancer Inst.* **104,** 1503–1513 (2012).
12. Le Gallo, M. *et al.* Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nature Genet.* **44,** 1310–1315 (2012).
13. Salvesen, H. B. *et al.* Integrated genomic profiling of endometrial carcinoma associates aggressive tumors with indicators of PI3 kinase activation. *Proc. Natl Acad. Sci. USA* **106,** 4834–4839 (2009).
14. Cowin, P. A. *et al.* LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin. *Cancer Res.* **72,** 4060–4073 (2012).
15. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
16. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150,** 1121–1134 (2012).
17. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
18. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genet.* **45,** 136–144 (2013).
19. Bartosch, C. *et al.* Endometrial carcinomas: a review emphasizing overlapping and distinctive morphological and immunohistochemical features. *Adv. Anat. Pathol.* **18,** 415–437 (2011).
20. Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Proc. Natl Acad. Sci. USA* **96,** 8681–8686 (1999).
21. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* **22,** 271–282 (2012).
22. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17,** 510–522 (2010).
23. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25,** 2906–2912 (2009).
24. Hockenbery, D., Nunez, G., Milliman, C., Schreiber, R. D. & Korsmeyer, S. J. Bcl-2 is an inner mitochondrial membrane protein that blocks programmed cell death. *Nature* **348,** 334–336 (1990).
25. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22,** 398–406 (2012).
26. Li, J., Mizukami, Y., Zhang, X., Jo, W. S. & Chung, D. C. Oncogenic K-*ras* stimulates Wnt signaling in colon cancer through inhibition of GSK-3β. *Gastroenterology* **128,** 1907–1918 (2005).
27. Zorn, A. M. *et al.* Regulation of Wnt signaling by Sox proteins: XSox17 α/β and XSox3 physically interact with β-catenin. *Mol. Cell* **4,** 487–498 (1999).
28. Sinner, D. *et al.* Sox17 and Sox4 differentially regulate β-catenin/T-cell factor activity and proliferation of colon carcinoma cells. *Mol. Cell. Biol.* **27,** 7802–7815 (2007).
29. Pollock, P. M. *et al.* Frequent activating FGFR2 mutations in endometrial carcinomas parallel germline mutations associated with craniosynostosis and skeletal dysplasia syndromes. *Oncogene* **26,** 7158–7162 (2007).
30. Fleming, G. F. *et al.* Phase II trial of trastuzumab in women with advanced or recurrent, HER2-positive endometrial carcinoma: a Gynecologic Oncology Group study. *Gynecol. Oncol.* **116,** 15–20 (2010).
31. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26,** i237–i245 (2010).
32. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
33. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
34. Clarke, B. A. & Gilks, C. B. Endometrial carcinoma: controversies in histopathological assessment of grade and tumour cell type. *J. Clin. Pathol.* **63,** 410–415 (2010).
35. Yemelyanova, A. *et al.* Utility of p16 expression for distinction of uterine serous carcinomas from endometrial endometrioid and endocervical adenocarcinomas: immunohistochemical analysis of 201 cases. *Am. J. Surg. Pathol.* **33,** 1504–1514 (2009).
36. Gilks, C. B., Oliva, E. & Soslow, R. A. Poor inter-observer reproducibility in the diagnosis of high-grade endometrial carcinoma. *Am. J. Surg. Pathol.* **91,** 248A (2012).
37. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
38. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11,** 367 (2010).
39. Houseman, E. A. *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9,** 365 (2008).
40. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101,** 4164–4169 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We wish to thank all patients and families who contributed to this study. We thank M. Sheth and L. Lund for administrative coordination of TCGA activities, G. Monemvasitis for editing the manuscript, and C. Gunter for critical reading of the manuscript. This work was supported by the following grants from the US National Institutes of Health: 5U24CA143799-04, 5U24CA143835-04, 5U24CA143840-04, 5U24CA143843-04, 5U24CA143845-04, 5U24CA143848-04, 5U24CA143858-04, 5U24CA143866-04, 5U24CA143867-04, 5U24CA143882-04, 5U24CA143883-04, 5U24CA144025-04, U54HG003067-11, U54HG003079-10 and U54HG003273-10.

Author Contributions The TCGA Research Network contributed collectively to this study. Biospecimens were provided by the tissue source sites and processed by the biospecimen core resource. Data generation and analyses were performed by the genome sequencing centres, cancer genome characterization centres and genome data analysis centres. All data were released through the data coordinating centre. The National Cancer Institute and National Human Genome Research Institute project teams coordinated project activities. We also acknowledge the following TCGA investigators who made substantial contributions to the project: N.S. (manuscript coordinator); J. Gao (data coordinator); C.K. and L. Ding (DNA sequence analysis); W.Z. and Y.L. (mRNA sequence analysis); H.S. and P.W.L. (DNA methylation analysis); A.D.C. and I.P. (copy number analysis); S.L. and A. Hadjipanayis (translocations); N.S., N.W. G.C., C.C.B. and C.Y. (pathway analysis); Andy C. and A.G.R. (miRNA sequence analysis); R. Broaddus, P.J.G., G.B.M. and R.A.S. (pathology and clinical expertise); G.B.M., H.L. and R.A. (reverse phase protein arrays); P.J.G. and R.B. (disease experts); G.B.M. and R.K. (manuscript editing); D.A.L. and E.R.M. (project chairs).

Author Information The primary and processed data used to generate the analyses presented here are deposited at the Data Coordinating Center (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp); all of the primary sequence files are deposited in CGHub (https://cghub.ucsc.edu/). Sample lists, data matrices and supporting data can be found at: (https://tcga-data.nci.nih.gov/docs/publications/ucec_2013/). The data can be explored via the cBio Cancer Genomics Portal (http://cbioportal.org). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A.L. (levine2@mskcc.org).

The Cancer Genome Atlas Research Network (Participants are arranged by area of contribution and then by institution.)

Genome sequencing centres: Broad Institute Gad Getz[1], Stacey B. Gabriel[1], Kristian Cibulskis[1], Eric Lander[1], Andrey Sivachenko[1], Carrie Sougnez[1], Mike Lawrence[1]; **Washington University in St Louis** Cyriac Kandoth[2], David Dooling[2], Robert Fulton[2], Lucinda Fulton[2], Joelle Kalicki-Veizer[2], Michael D. McLellan[2], Michelle O'Laughlin[2], Heather Schmidt[2], Richard K. Wilson[2], Kai Ye[2], Li Ding[2], Elaine R. Mardis[2]

Genome characterization centres: British Columbia Cancer Agency Adrian Ally[3], Miruna Balasundaram[3], Inanc Birol[3], Yaron S. N. Butterfield[3], Rebecca Carlsen[3], Candace Carter[3], Andy Chu[3], Eric Chuah[3], Hye-Jung E. Chun[3], Noreen Dhalla[3], Ranabir Guin[3], Carrie Hirst[3], Robert A. Holt[3], Steven J. M. Jones[3], Darlene Lee[3], Haiyan I. Li[3], Marco A. Marra[3], Michael Mayo[3], Richard A. Moore[3], Andrew J. Mungall[3], Patrick Plettner[3], Jacqueline E. Schein[3], Payal Sipahimalani[3], Angela Tam[3], Richard J. Varhol[3], A. Gordon Robertson[3]; **Broad Institute** Andrew D. Cherniack[1], Itai Pashtan[1,4,5], Gordon Saksena[1], Robert C. Onofrio[1], Steven E. Schumacher[1], Barbara Tabak[1], Scott L. Carter[1], Bryan Hernandez[1], Jeff Gentry[1], Helga B. Salvesen[1,6,7], Kristin Ardlie[1], Gad Getz[1], Wendy Winckler[1], Rameen Beroukhim[1,8], Stacey B. Gabriel[1], Matthew Meyerson[1,8]; **Harvard Medical School/Brigham & Women's Hospital/MD Anderson Cancer Center** Angela Hadjipanayis[9], Semin Lee[10], Harshad S. Mahadeshwar[11], Peter Park[10,12], Alexei Protopopov[11], Xiaojia Ren[9], Sahil Seth[11], Xingzhi Song[11], Jiabin Tang[11], Ruibin Xi[10], Lixing Yang[10], Dong Zeng[11], Raju Kucherlapati[9], Lynda Chin[1,11], Jianhua Zhang[11]; **University of North Carolina** J. Todd Auman[13,14], Saianand Balu[15], Tom Bodenheimer[15], Elizabeth Buda[15], D. Neil Hayes[15,16], Alan P. Hoyle[15], Stuart R. Jefferys[15], Corbin D. Jones[17], Shaowu Meng[15], Piotr A. Mieczkowski[18], Lisle E. Mose[15], Joel S. Parker[15], Charles M. Perou[15,18,19], Jeff Roach[20], Yan Shi[15], Janae V. Simons[15], Mathew G. Soloway[15], Donghui Tan[15], Michael D. Topal[15,19], Scot Waring[15], Junyuan Wu[15], Katherine A. Hoadley[15,18]; **University of Southern California & Johns Hopkins** Stephen B. Baylin[21], Moiz S. Bootwalla[22], Phillip H. Lai[22], Timothy J. Triche Jr[22], David J. Van Den Berg[22], Daniel J. Weisenberger[22], Peter W. Laird[22], Hui Shen[22]

Genome data analysis centres: Broad Institute Lynda Chin[1,11], Jianhua Zhang[11], Gad Getz[1], Juok Cho[1], Daniel DiCara[1], Scott Frazer[1], David Heiman[1], Rui Jing[1], Pei Lin[1], Will Mallard[1], Petar Stojanov[1], Doug Voet[1], Hailei Zhang[1], Lihua Zou[1], Michael Noble[1], Mike

Lawrence[1]; **Institute for Systems Biology** Sheila M. Reynolds[23], Ilya Shmulevich[23]; **Memorial Sloan-Kettering Cancer Center** B. Arman Aksoy[24], Yevgeniy Antipin[24], Giovanni Ciriello[24], Gideon Dresdner[24], Jianjiong Gao[24], Benjamin Gross[24], Anders Jacobsen[24], Marc Ladanyi[25], Boris Reva[24], Chris Sander[24], Rileen Sinha[24], S. Onur Sumer[24], Barry S. Taylor[26], Ethan Cerami[24], Nils Weinhold[24], Nikolaus Schultz[24], Ronglai Shen[27]; **University of California, Santa Cruz/Buck Institute** Stephen Benz[28], Ted Goldstein[28], David Haussler[28], Sam Ng[28], Christopher Szeto[28], Joshua Stuart[28], Christopher C. Benz[29], Christina Yau[29]; **The University of Texas MD Anderson Cancer Center** Wei Zhang[30,31], Matti Annala[30,31,32], Bradley M. Broom[33], Tod D. Casasent[33], Zhenlin Ju[33], Han Liang[33], Guoyan Liu[30,31], Yiling Lu[34], Anna K. Unruh[33], Chris Wakefield[33], John N. Weinstein[33], Nianxiang Zhang[33], Yuexin Liu[30,31], Russell Broaddus[31], Rehan Akbani[33], Gordon B. Mills[34]

Biospecimen core resource: Nationwide Children's Hospital Christopher Adams[35], Thomas Barr[35], Aaron D. Black[35], Jay Bowen[35], John Deardurff[35], Jessica Frick[35], Julie M. Gastier-Foster[35,36], Thomas Grossman[35], Hollie A. Harper[35], Melissa Hart-Kothari[35], Carmen Helsel[35], Aaron Hobensack[35], Harkness Kuck[35], Kelley Kneile[35], Kristen M. Leraas[35], Tara M. Lichtenberg[35], Cynthia McAllister[35], Robert E. Pyatt[35], Nilsa C. Ramirez[35,36], Teresa R. Tabler[35], Nathan Vanhoose[35], Peter White[35], Lisa Wise[35], Erik Zmuda[35]

Tissue source sites: Asterand Nandita Barnabas[37], Charlenia Berry-Green[37], Victoria Blanc[37], Lori Boice[38], Michael Button[37], Adam Farkas[37], Alex Green[37], Jean MacKenzie[37], Dana Nicholson[37]; **British Columbia Cancer Agency** Steve E. Kalloger[39,40], C. Blake Gilks[39,40]; **Cedars-Sinai Medical Center** Beth Y. Karlan[41], Jenny Lester[41], Sandra Orsulic[41]; **Christiana Care** Mark Borowsky[42], Mark Cadungog[42], Christine Czerwinski[42], Lori Huelsenbeck-Dill[42], Mary Iacocca[42], Nicholas Petrelli[42], Brenda Rabeno[42], Gary Witkin[42]; **Cureline** Elena Nemirovich-Danchenko[43], Olga Potapova[43], Daniil Rotin[43]; **Duke University** Andrew Berchuck[44]; **Gynecologic Oncology Group** Michael Birrer[45], Phillip DiSaia[46], Laura Monovich[47]; **International Genomics Consortium** Erin Curley[48], Johanna Gardner[48], David Mallery[48], Robert Penny[48]; **Mayo Clinic** Sean C. Dowdy[49], Boris Winterhoff[49], Linda Dao[50], Bobbie Gostout[49], Alexandra Meuter[49], Attila Teoman[49]; **Memorial Sloan-Kettering Cancer Center** Fanny Dao[51], Narciso Olvera[51], Faina Bogomolniy[51], Karuna Garg[52], Robert A. Soslow[52], Douglas A. Levine[51]; **N. N. Blokhin Russian Cancer Research Center** Mikhail Abramov[53]; **Ontario Tumour Bank** John M. S. Bartlett[54], Sugy Kodeeswaran[54], Jeremy Parfitt[55]; **St Petersburg Academic University** Fedor Moiseenko[56]; **University Health Network** Blaise A. Clarke[57]; **University of Hawaii** Marc T. Goodman[58,59], Michael E. Carney[58], Rayna K. Matsuno[58]; **University of North Carolina** Jennifer Fisher[38], Mei Huang[38], W. Kimryn Rathmell[15], Leigh Thorne[38], Linda Van Le[38]; **University of Pittsburgh** Rajiv Dhir[60], Robert Edwards[60], Esther Elishaev[60], Kristin Zorn[60]; **The University of Texas MD Anderson Cancer Center** Russell Broaddus[31]; **Washington University School of Medicine** Paul J. Goodfellow[36,61], David Mutch[61]

Disease analysis working group: Nikolaus Schultz[24], Yuexin Liu[30,31], Rehan Akbani[33], Andrew D. Cherniack[1], Ethan Cerami[24], Nils Weinhold[24], Hui Shen[22], Katherine A. Hoadley[15,18], Ari B. Kahn[62], Daphne W. Bell[63], Pamela M. Pollock[64], Chen Wang[65], David A.Wheeler[66], Eve Shinbrot[66], Beth Y. Karlan[41], Andrew Berchuck[44], Sean C. Dowdy[49], Boris Winterhoff[49], Marc T. Goodman[58,59], A. Gordon Robertson[3], Rameen Beroukhim[1,8], Itai Pashtan[1,4,5], Helga B. Salvesen[1,6,7], Peter W. Laird[22], Michael Noble[1], Joshua Stuart[28], Li Ding[2], Cyriac Kandoth[2], C. Blake Gilks[39,40], Robert A. Soslow[52], Paul J. Goodfellow[36,61], David Mutch[61], Russell Broaddus[31], Wei Zhang[30,31], Gordon B. Mills[34], Raju Kucherlapati[9], Elaine R. Mardis[2], Douglas A. Levine[51]

Data coordination centre: Brenda Ayala[62], Anna L. Chu[62], Mark A. Jensen[62], Prachi Kothiyal[62], Todd D. Pihl[62], Joan Pontius[62], David A. Pot[62], Eric E. Snyder[62], Deepak Srinivasan[62], Ari B. Kahn[62]

Project team: National Cancer Institute Kenna R. Mills Shaw[67], Margi Sheth[67], Tanja Davidsen[67], Greg Eley[68]; Martin L. Ferguson[69], John A. Demchok[67], Liming Yang[67]; **National Human Genome Research Institute** Mark S. Guyer[70], Bradley A. Ozenberger[70], Heidi J. Sofia[70]

Writing committee: Cyriac Kandoth[2], Nikolaus Schultz[24], Andrew D. Cherniack[1], Rehan Akbani[33], Yuexin Liu[30,31], Hui Shen[22], A. Gordon Robertson[3], Itai Pashtan[1,4,5], Ronglai Shen[27], Christopher C. Benz[29], Christina Yau[29], Peter W. Laird[22], Li Ding[2], Wei Zhang[30,31], Gordon B. Mills[34], Raju Kucherlapati[9], Elaine R. Mardis[2] & Douglas A. Levine[51]

[1]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University Cambridge, Massachusetts 02142, USA. [2]The Genome Institute, Washington University, St Louis, Missouri 63108, USA. [3]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z, Canada. [4]Department of Radiation Oncology, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [5]Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [6]Department of Obstetrics and Gynecology, Haukeland University Hospital, 5021 Bergen, Norway. [7]Department of Clinical Medicine, University of Bergen, 5020 Bergen, Norway. [8]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [9]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [10]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA. [11]Institute for Applied Cancer Science, Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, Texas 77054, USA. [12]Informatics Program, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [13]Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [14]Institute for Pharmacogenetics and Individualized Therapy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [15]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [16]Department of Internal Medicine, Division of Medical Oncology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [17]Department of Biology, University of North Carolina at Chapel Hill, North Carolina 27599, USA. [18]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [19]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [20]Research Computing Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [21]Cancer Biology Division, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, Maryland 21231, USA. [22]University of Southern California Epigenome Center, University of Southern California, Los Angeles, California 90089, USA. [23]Institute for Systems Biology, Seattle, Washington 98109, USA. [24]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [25]Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [26]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California 94158, USA. [27]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [28]Department of Biomolecular Engineering and Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. [29]Buck Institute for Age Research, Novato, California 94945, USA. [30]Cancer Genomics Core Laboratory, University of Texas MD Anderson Cancer Center, Houston, Texas 77054, USA. [31]Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [32]Tampere University of Technology Korkeakoulunkatu 10, FI-33720 Tampere, Finland. [33]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [34]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [35]The Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, USA. [36]The Ohio State University, Columbus, Ohio 43210, USA. [37]Asterand, Detroit, Michigan 48202, USA. [38]University of North Carolina, Chapel Hill, North Carolina 27599, USA. [39]OvCaRe British Columbia, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada. [40]Department of Pathology & Laboratory Medicine, The University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. [41]Women's Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, USA. [42]Helen F Graham Cancer Center at Christiana Care, Newark, Delaware 19713, USA. [43]Cureline, Inc., South San Francisco, California 94080, USA. [44]Duke University Medical Center, Duke Cancer Institute, Durham, North Carolina 27710, USA. [45]Harvard Medical School, Massachusetts General Hospital Cancer Center, Boston, Massachusetts 02114, USA. [46]University of California Medical Center, Irvine, Orange California 92868, USA. [47]GOG Tissue Bank, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, USA. [48]International Genomics Consortium, Phoenix, Arizona 85004, USA. [49]Department of OB Gyn, Division of Gynecologic Oncology, Mayo Clinic, Rochester, Minnesota 55905, USA. [50]Department of Pathology, Mayo Clinic, Rochester, Minnesota 55905, USA. [51]Gynecology Service, Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [52]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [53]N. N. Blokhin Russian Cancer Research Center RAMS, Moscow 115478, Russia. [54]Ontario Tumour Bank, Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada. [55]Ontario Tumour Bank, London Health Sciences Centre, London, Ontario N6A 5A5, Canada. [56]St Petersburg Academic University, St Petersburg 199034, Russia. [57]Department of Pathology, University Health Network, Toronto, Ontario M5G 2C4, Canada. [58]University of Hawaii, Honolulu, Hawaii 96813, USA. [59]Cedars-Sinai Medical Center, Los Angeles, California 90024, USA. [60]University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. [61]Washington University School of Medicine, St Louis, Missouri 63110, USA. [62]SRA International, Fairfax, Virgina 22033, USA. [63]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [64]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane 4059, Australia. [65]Department of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota 55905, USA. [66]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [67]The Cancer Genome Atlas Program Office, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [68]Scimentis, LLC, Atlanta, Georgia 30666, USA. [69]MLF Consulting, Arlington, Maryland 02474, USA. [70]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

doi:10.1038/nature12112

Hippocampal place-cell sequences depict future paths to remembered goals

Brad E. Pfeiffer[1] & David J. Foster[1]

Effective navigation requires planning extended routes to remembered goal locations. Hippocampal place cells have been proposed to have a role in navigational planning, but direct evidence has been lacking. Here we show that before goal-directed navigation in an open arena, the rat hippocampus generates brief sequences encoding spatial trajectories strongly biased to progress from the subject's current location to a known goal location. These sequences predict immediate future behaviour, even in cases in which the specific combination of start and goal locations is novel. These results indicate that hippocampal sequence events characterized previously in linearly constrained environments as 'replay' are also capable of supporting a goal-directed, trajectory-finding mechanism, which identifies important places and relevant behavioural paths, at specific times when memory retrieval is required, and in a manner that could be used to control subsequent navigational behaviour.

A fundamental purpose of memory lies in using previous experience to inform current choices, directing behaviour towards reward and away from negative consequences based on knowledge of prior outcomes in similar situations. Goal-directed spatial navigation—planning extended routes to remembered locations—requires both memory of the goal location and knowledge of the intervening terrain to determine an efficient and safe path. The hippocampus has long been known to have a critical role in spatial memory[1,2] and memory for events[3,4], and it has been proposed that the hippocampus may have a fundamental role in calculating routes to goals, especially under conditions demanding behavioural flexibility[1,5–8]. This proposal stems largely from the discovery that excitatory neurons of the hippocampus show spatially localized place responses during exploration[1]. However, it has been a challenge to understand how individual place responses tied to the current location might be informative about other locations that the animal cares about, such as the remembered goal[9], or the set of locations defining a route[10,11].

Techniques to record simultaneously from multiple hippocampal place cells[12] have been used to show that place cells systematically represent positions other than the current location. The early discovery of phase precession of place-cell spikes relative to theta frequency oscillations in the local field potential[13] led to the hypothesis that place cells fire in sequences within a theta cycle, and thus represent places behind or ahead of the animal[14–16]. Theta sequences have since been demonstrated experimentally across place-cell populations[17]. Also during theta, place-cell activity seems to 'sweep' ahead of an animal located at a choice point[18], leading to the hypothesis that such activity could support the evaluation of alternatives during decision making[19]. A separate group of phenomena termed 'replay' has been found during sleep[20,21] and non-exploratory awake periods[22], and is associated with sharp-wave-ripple (SWR) events in the hippocampal local field potential (with the sole exception of replay during rapid eye movement sleep[20]). In replay, simultaneously recorded populations of place cells show reactivation of temporal sequences reflecting prior behavioural trajectories up to 10-m long[23]. Although these forms of non-local activity are now well established[17,23–26], it has proven difficult to establish a predictive relationship between non-local place-cell activity and behaviour[18,26], because of the twofold technical problem of ensuring adequate behavioural sampling of the environment while recording from sufficient numbers of place cells. Thus it remains unknown whether non-local place-cell activity can specify remembered goals, or define specific routes that the animal will take.

Depiction of two-dimensional trajectories

We recorded from hippocampal neurons while rats performed a spatial memory task, using the statistical power of an open-field design in which the goal was one of 36 clearly separated locations within a $2\,\mathrm{m} \times 2\,\mathrm{m}$ arena (Fig. 1a). We addressed the sampling problem by combining random foraging and goal-directed behaviour, and by implanting miniaturized lightweight microdrives supporting 40 independently adjustable tetrodes, with 20 tetrodes targeted to each dorsal hippocampal area CA1 (Supplementary Fig. 1), to record simultaneously from up to 250 hippocampal neurons with well-defined place fields. Our task, incorporating elements from previous task designs[9,27–29], was composed of trials each consisting of two phases: in phase one, the rat was required to forage to obtain reward (liquid chocolate) in an unknown location (Random). In phase two, the rat could obtain reward in a predictable reward location (Home). The transition to the next phase or trial was automatic upon consummation of the reward, and was not signalled to the animal. The task incorporated several features. First, because the shortest routes in phase one and two were matched, it was determined that animals could remember Home, but could not detect Random locations, because latencies and path lengths were significantly shorter for Home-bound trajectories (Fig. 1b–d). Second, the Home location was moved to a new location each day. Thus, animals were required to learn a new goal location, demanding a flexible behavioural response that was more likely to engage the hippocampus than a fixed reference-memory response[27,30,31]. Third, for the first 19 trials of each day, the Random locations were non-repeating. Hence during this period, every Home-bound trajectory was always a novel combination of current location and goal location. Thus, our task probed both memory for the goal location and flexible planning of a novel route to get there.

We implanted four well-trained rat subjects with the 40-tetrode microdrive for electrophysiological recording. Large numbers of

[1]Solomon H Snyder Department of Neuroscience, Johns Hopkins University School Of Medicine, Baltimore, Maryland 21205, USA.

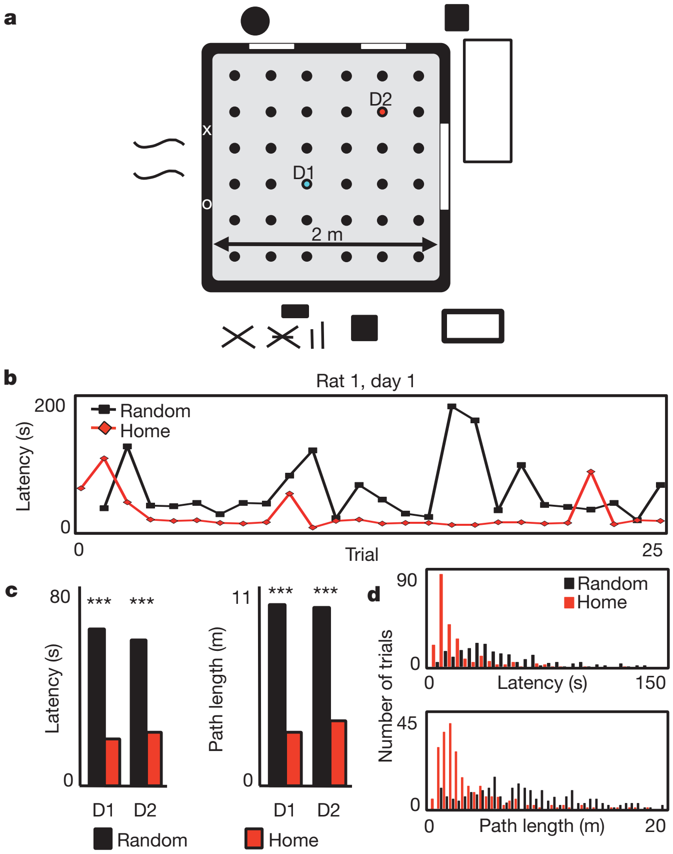


Figure 1 | Behaviour in the open-field spatial memory task. a, Schema of arena and room, reward wells (circles), and Home location for days 1 and 2 (D1, cyan; D2, red). **b**, Per-trial latency to reach Home or Random well location for rat 1 (R1) on D1. **c**, Mean latency and path length to reach Home or Random well location across all rats for D1 and D2. *P*-values (Wilcoxon rank-sum test): latency D1 5.5×10^{-19}, D2 9.7×10^{-14}; path D1 2.7×10^{-19}, D2 5.2×10^{-16}. **d**, Histogram of latencies (5-s bins) and path lengths (50-cm bins) for all trials (shown to 150 s and 20 m, respectively) *P*-values (Kolmogorov–Smirnov test): latency 2.6×10^{-2}; path 9.1×10^{-4}.

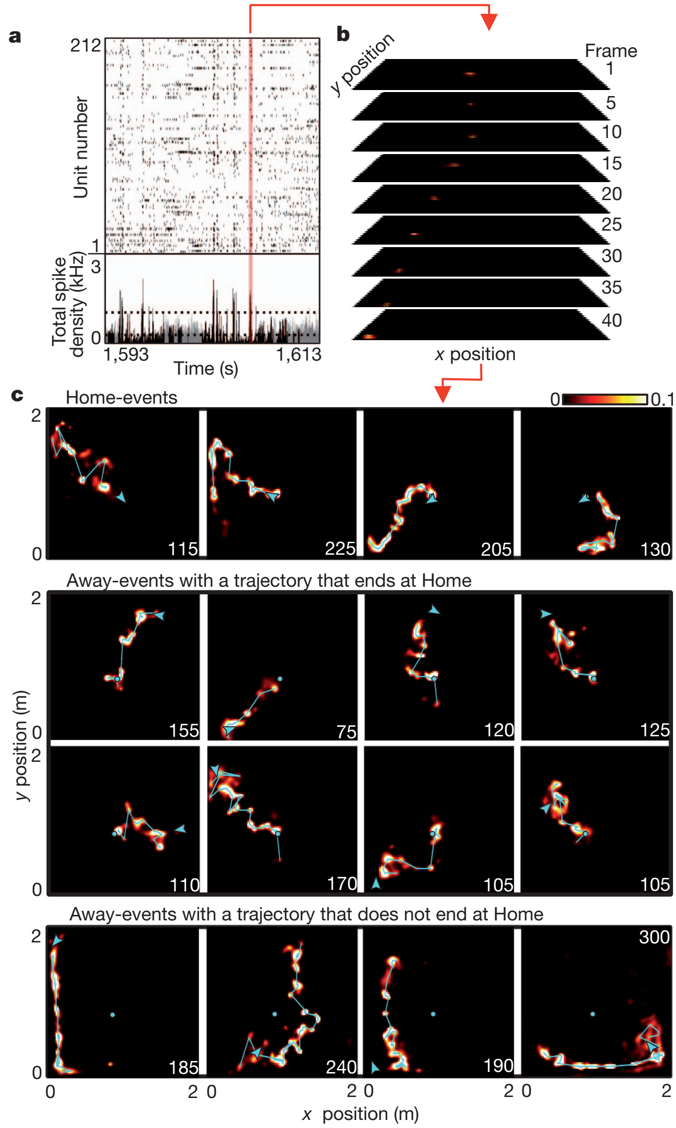


Figure 2 | Trajectory events. a, Raster plot (top) and spike density (bottom) of simultaneous unit activity for R1,D1 for representative epoch. Periods of immobility denoted in black. Dashed lines represent candidate event detection threshold. **b**, Position posterior probabilities in selected frames for the candidate event in **a**. **c**, 16 representative events (of 274) for R1,D1, decoded and summed across time. Values indicated by colour bar. Event duration (in ms) in right corner. Cyan circle, Home well. Cyan line, peak probability for each timeframe. Cyan arrowhead, position and head direction of rat at time of event. Videos of each event available in online Supplementary Video 2.

well-isolated units (Supplementary Fig. 2)[32] were recorded simultaneously during behavioural sessions on two consecutive days (212 and 250 units active during exploration from rat 1 on experimental days 1 and 2, respectively; 166 and 193 units from rat 2 on days 1 and 2; 133 and 106 units from rat 3 on days 1 and 2; 103 and 175 units from rat 4 on days 1 and 2). The recorded units demonstrated position-specific firing patterns ('place fields') that were distributed throughout the environment (Supplementary Figs 3–5), and a memory-less, uniform prior Bayesian decoding algorithm[23] allowed us to estimate the spatial location of the rat accurately from the recorded spike trains throughout the experiment (Supplementary Fig. 6 and Supplementary Video 1). We identified candidate events as brief increases in population spiking activity during periods of immobility while the rat performed the task (Fig. 2a) and applied the decoding algorithm to the population spike trains (Fig. 2b). During many candidate events, decoded position revealed temporally compressed, two-dimensional trajectories across the environment (Fig. 2c and Supplementary Video 2). We applied length, duration and smoothness criteria to the decoded positions of candidate events to define 'trajectory events' (see Methods). We found between 144 and 373 trajectory events per session (between 25.3% and 43.9% of candidate events) with a mean duration of 103.6 ms, and path lengths that ranged from 40.0 cm to 199.1 cm (Supplementary Fig. 7 and Supplementary Table 1). We tested the probability that trajectory events could have occurred by chance, using two separate Monte-Carlo shuffle methods which varied either cell identity or place field position (see Methods). Zero (out of 2,028) trajectory events had a *P*-value greater than 0.02 under either method, indicating that all trajectory events were statistically significant events. Spectrogram analysis of trajectory events strongly matched SWR events identified within the same experimental sessions (Supplementary Fig. 8a). In addition, an overwhelming majority of trajectory events were coincident with SWR events (Supplementary Fig. 8b). Theta power, which is high during exploration, was significantly decreased immediately before and after trajectory events (Supplementary Fig. 8c). Collectively, these data indicate that trajectory events are functionally similar to the SWR-associated events previously reported on linear tracks as 'replay'[21–26].

Trajectory events over-represent the goal

To examine whether non-local spatial information present in trajectory events contributes to or is affected by acquisition or expression of

a spatial memory (the novel Home location), we divided the observed trajectory events into those that were initiated while the rat was at the Home location ('home-events') and those that were initiated while the rat was elsewhere ('away-events'). There was no difference in the rate of occurrence of sharp-wave/ripple events or of trajectory events between Home and Random locations (Supplementary Figs 9 and 10). As expected, home-events showed strong representation of the Home location (Fig. 3a, c and Fig. 2c, top row), probably owing to initiation bias, a tendency for hippocampal events to reflect a path that begins at the rat's current location[22–24] (but see refs 25, 26). Strikingly, we observed that away-events also showed an increased representation of the Home location (Fig. 3b, d and Supplementary Fig. 11), a finding that cannot be explained through initiation bias. Consistent with this observation, many away-events depicted a trajectory that ended at Home (Fig. 2c, middle rows; Supplementary Videos 3–7). Quantification confirmed that the Home location was significantly over-represented in away-events relative to other locations on the open field (Fig. 3e, left; Supplementary Fig. 12) and that away-events were more likely to end their trajectories at the Home location than any other region of the arena (Fig. 3e, centre).

Importantly, the region of increased representation changed accordingly when the location of the Home well was moved on experimental day 2. The heightened representation of Home in away-events was present even when the analysis was restricted to the first 19 trials, when the specific Random–Home combinations were novel (Supplementary Fig. 13). The increased representation of Home in away-events was not a simple function of increased familiarity with or time spent at the Home location, as other regions of the arena with greater occupancy times did not show strong representations in trajectory events (Fig. 3e, right). The overexpression of the Home location in away-events could not be accounted for by either occupancy time or the spatial distribution of place fields (Supplementary Figs 14 and 15). Further, when we restricted our analysis to vectorized trajectories rather than entire posterior probabilities, the Home location remained over-represented in away-events (Supplementary Fig. 16). Thus, trajectory events in the hippocampus over-represent a known goal location in a manner which cannot be explained solely by occupancy time or place-field representation.

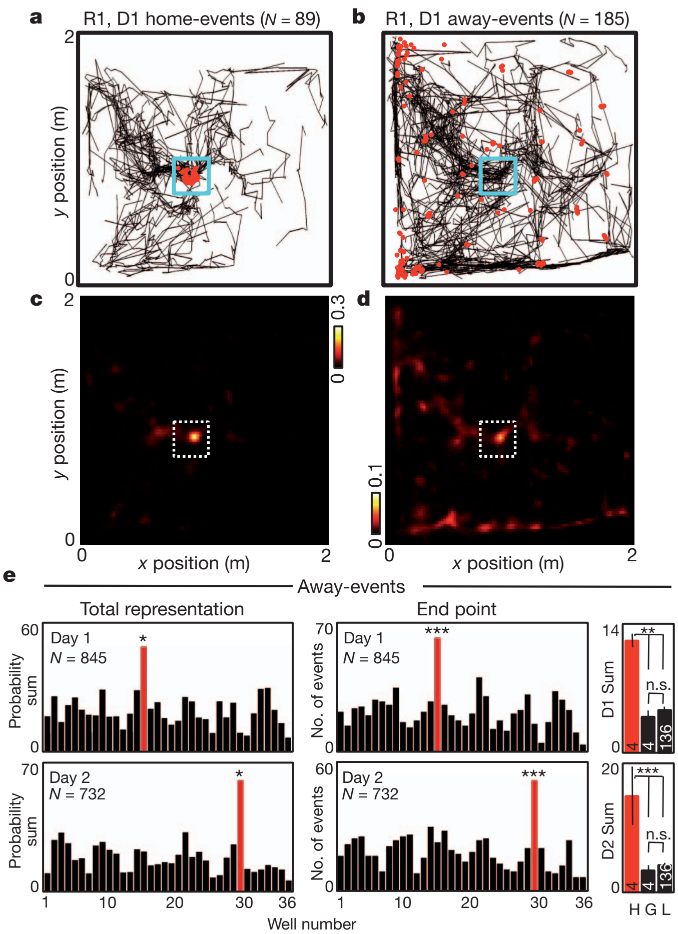


Figure 3 | Remote representation of goal location. a–d, Vectorized trajectories (**a**, **b**) and average posterior probability sum (**c**, **d**) of all confirmed home-events (left) and away-events (right) for R1,D1. Red dots in **a**, **b**, rat location at time of event. Dashed box in **c**, **d**, Home location. **e**, Left, posterior probability sum for all away-events across all rats. Home (red) is a statistical outlier. *P*-value (Grubbs' test for outliers): *D1* 2.3×10^{-2} (Lilliefors test, *P*-value 0.15); D2 1.1×10^{-2} (Lilliefors *P*-value 0.32). Centre, number of away-events across all rats in which the final frame peak posterior probability was at each well. Home (red) is a statistical outlier. *P*-values (Grubbs' test for outliers): D1 6.9×10^{-4} (Lilliefors test, *P*-value 0.29); D2 6.0×10^{-4} (Lilliefors *P*-value 0.42). Right, as left, but mean $\pm$ s.e.m. for Home (H), all wells with greater in-session total occupancy than Home (G), and all wells with less occupancy than Home (L). *P*-values (ANOVA, Tukey–Kramer post-hoc multiple comparison): D1 H vs G 2.9×10^{-3}, H vs L 8.5×10^{-5}, G vs L 0.91; D2 H vs G 7.4×10^{-8}, H vs L $< 1 \times 10^{-10}$, G vs L 0.82.

No over-representation of non-goals

We proposed that the over-representation of locations in trajectory events was selective for behaviourally relevant locations. The task was designed so that the previous Random well was never a correct behavioural goal, and so we proposed in particular that the previous Random well would not be over-represented in trajectory events. To equalize comparison between away-events and home-events, we rotated and scaled all home-events such that the distance and direction from the rat's physical location at the time of each event to the previously active Random location was the same across all home-events (Fig. 4a). Similarly, we rotated and scaled all away-events according to the direction and distance to the Home location (Fig. 4b), and as a control we rotated and scaled all home-events according to the direction and distance to the immediately future (but unknown and not yet baited) Random well location. All rotated/scaled trajectory events showed a strong representation of the rat's physical location (Fig. 4c–e) due to initiation bias. However, whereas the rotated/scaled away-events showed a strong representation of the Home location (Fig. 4d), rotated home-events showed little representation of the previously active (Fig. 4c) or immediately-to-be active (Fig. 4e) Random locations. Indeed, we observed a significant decrease in the representation of the previous Random location in home-events compared to the representation of the Home location in away-events (Fig. 4f). These data show that hippocampal trajectory events reflect the demands of the task by selectively over-representing the immediately relevant Home location and not the irrelevant previous Random location.

Trajectory events reflect future behavioural path

The initiation and termination bias that we observed suggested that away events depict the future trajectory to Home, indicative of a planning mechanism to guide behaviour. To test this hypothesis, we quantified the correspondence between trajectory events and the behavioural path in the immediate future, or immediate past (Fig. 5a, b and Supplementary Fig. 17). We calculated the angular displacement between trajectory and path at progressively increasing radii from the current location (Fig. 5a, b). Away-events were strongly concentrated around zero angular displacement assessed against the future path, and more broadly distributed with respect to the past path (Fig. 5c), and this difference was verified in terms of the mean absolute angular displacement for each event (Fig. 5d). Home events showed a weaker representation of future path, and an apparent anti-correlation with past path, which might have reflected the fact that the path back to the previous Random well was never correct (Fig. 5e, f). Away-events were significantly closer to the rat's future path than were home-events

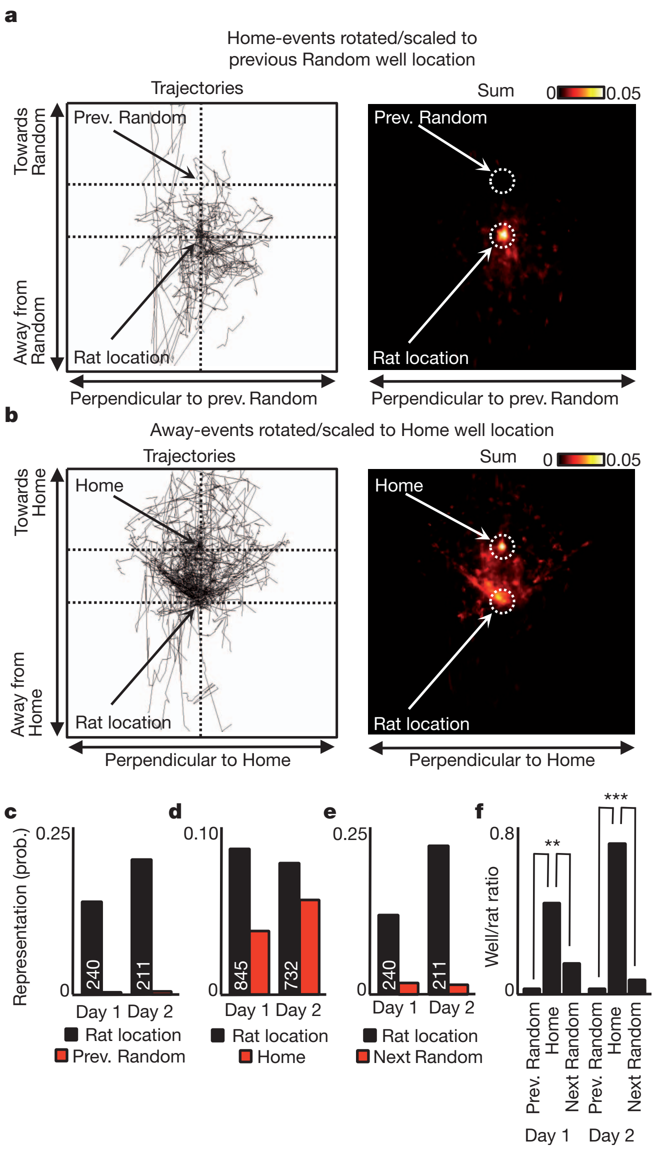


Figure 4 | Representation of relevant vs. irrelevant locations. **a**, Vectorized trajectories (left) and average posterior probability sum (right) of all home-events for R1,D1, centred by the rat's physical location at time of event and rotated and scaled according to direction and distance to the previously rewarded Random location. White circles, quantified regions. Prev., previous. **b**, As **a**, for Home. **c–e**, Across all rats, mean representation of quantified regions as in **a**, **b**. Event number displayed on bar. **f**, Normalized ratio of well/rat representation for **c–e**. *P*-values (Wilcoxon rank-sum test): D1 Home vs prev. Random 4.4×10^{-16}, Home vs next Random 9.9×10^{-3}; D2 Home vs prev. Random 3.1×10^{-20}, Home vs next Random 1.3×10^{-13}.

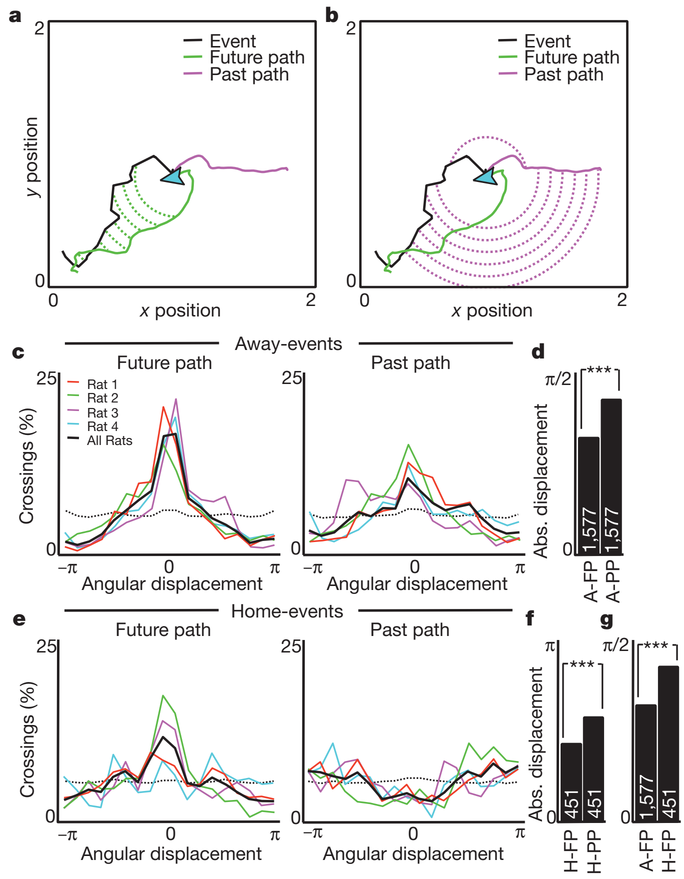


Figure 5 | Correspondence to past or future path. **a**, **b**, Representative event from R1,D1, demonstrating trajectory event vector (black), immediate future (green) and past (magenta) path (up to 10 s or 50 cm, whichever is greater), and angular displacement along the minor arc between event and future (**a**) or past (**b**) path at each crossing. **c**, Per cent of crossings across all events as a function of angular displacement for all away-events compared to future (left) or past (right) path. Dashed line indicates chance based upon 2,000 shuffled events. **d**, Mean absolute angular displacement for away-events compared to future (A-FP) or past (A-PP) path. Abs., absolute. **e**, **f**, As **c**, **d**, for home-events. **g**, Mean absolute angular displacement for future path for all away-events (A-FP) or home-events (H-FP). *P*-values (Wilcoxon rank-sum test): 8.60×10^{-31} (**d**); 3.54×10^{-17} (**f**); 7.25×10^{-16} (**g**).

(Fig. 5g), consistent with the goal-directed nature of Random-to-Home navigation. We conducted two further analyses of path correspondence, one based on the orientation of the depicted trajectory to a location occupied 10 s in the future or the past (Supplementary Fig. 18), and one based on the spatial overlap between smoothed versions of the trajectory and future or past path (Supplementary Fig. 19), with matching results. Rats showed no bias to face the direction of their immediately future path or the Home well location during away-events (Supplementary Fig. 20a, b). Furthermore, away-events were more spatially correlated with the rat's future path than with his current heading (Supplementary Fig. 20d–g). Thus, the strong reflection of the rat's future path in away-events could not be trivially explained as a representation of paths 'in front' of the rat, but rather suggested a more precise path-finding mechanism.

A flexible planning mechanism

If trajectory events reflect behavioural planning generally, they might also have depicted future behaviours when the animal did not proceed immediately to the Home location. Indeed, away-events closely matched the rat's future path regardless of whether the rat's future path took it to the Home location or elsewhere in the arena (Fig. 6a, c). For both cases, trajectories matched the future path more than the past path (Fig. 6b, d). We proposed that if trajectory events reflected an active process that could switch between goals, then before non-Home-seeking behaviours, not only would the representation of the non-Home-seeking path be enhanced, but the representation of the Home well would be reduced. Indeed, we found reduced Home representation in non-Home-seeking away-events compared to Home-seeking away-events (Fig. 6e).

We finally proposed that a flexible planning mechanism should be able to specify paths of novel importance (a novel combination of start and end points) over familiar terrain. The animals' behaviour showed evidence of this ability over the first 19 trials of each day. We therefore examined trajectory events during this period of each session. Away-events during this novel period also bore a strong match to the rat's future path (Fig. 6f and Supplementary Videos 3–7), and were closer to the rat's future path than its past path (Fig. 6g).

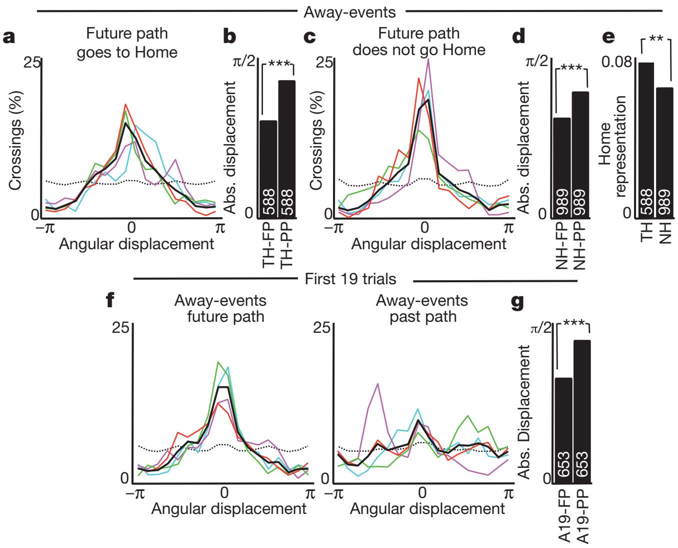


Figure 6 | Goal switching and flexibility in trajectory events. a, b, As Fig. 5c (left) and 5d, for away-events preceding behaviour ending at or crossing Home (future path, TH-FP; past path, TH-PP). **c, d**, As **a, b**, for away-events preceding behaviours directed elsewhere (future path, NH-FP; past path, NH-PP). **e**, Mean posterior probability representation of Home for same division of away-events (to Home, TH; not to Home, NH). **f, g**, As Fig. 5c, d, for away-events from the first 19 trials of each session (future path, A19-FP; past path, A19-PP). *P*-values (Wilcoxon rank-sum test): 4.96×10^{-22} (**b**); 1.12×10^{-13} (**d**); 9.60×10^{-3} (**e**);

Discussion

We have demonstrated that hippocampal SWR-associated trajectory events predict immediate future navigational behaviour. This finding follows a succession of results[8] reporting that SWR-associated sequences occur robustly during the awake state[22–26], that sequences are not always facsimiles of previous behavioural episodes[22–24,26,33] and can even depict novel combinations of previous experiences[26], and that sequences can be selective to the extent of not always reporting the most recent experience[26], or even necessarily experiences from the current environment[25]. Moreover, disruption studies using electrical stimulation contingent on SWR detection have revealed a role for sleep SWRs in learning[34,35], and a specific role for awake SWRs in working memory but not reference memory[36], which accords with the flexibility of trajectory events in response to a daily changing goal location[27,30,31]. Regarding our observation of stronger prediction before goal-finding than random foraging, it is likely that during the latter behaviour, an animal repeatedly makes online changes to his planned navigational trajectory, which would reduce its initial predictability. This strategic variability may be reflected in the over-dispersion of place-cell firing rates during random foraging[28,37]. Regarding the mechanism generating trajectory events, low-level mechanisms might have contributed, such as the spatial distributions of place cells' firing rates, although these did not account for the precise depiction of the goal location. Alternatively, it is equally possible that the spatial distributions of firing rates emerged as a consequence of the trajectory events. Simple models of encoding routes via direct experience cannot easily explain either the trial-by-trial switching of trajectory events between different goals (Home-seeking versus non-Home-seeking), or the trajectory events corresponding to novel Random–Home combinations[6,38,39], although the incorporation of contextual coding for the goal might account for some of this functionality[5,40]. It remains unknown whether trajectory events can reflect the calculation of optimal paths in more challenging navigational tasks that incorporate barriers to movement[41,42]. Finally, we might speculate on how the planning function of trajectory events operates. Trajectory depiction by place cells before behaviour might support a plasticity mechanism that reinforces the particular path, in a way that can be accessed locally during behaviour[43]. For example, trajectory events might drive associations between places en route and estimates of value[19,31,44,45] or chosen action[44,46] that could be accessed subsequently by local place-cell activation during goal-directed behaviour, perhaps in combination with a local look-ahead mechanism such as theta sequences.

In summary, our data reveal a flexible, goal-directed mechanism for the manipulation of previously acquired memories, in which behavioural trajectories to a remembered goal are depicted in the brain immediately before movement. Such findings address longstanding questions about the role of place cells in navigational learning and planning, as well as broader questions regarding the recall and use of stored memory. In particular, trajectory events relate to hippocampal function in multiple conceptual contexts: as a cognitive map in which routes to goals might be explored flexibly before behaviour[1], as an episodic memory system engaging in what has been termed 'mental time travel'[47], and as a substrate for the recall of imaginary events[48,49]. These conceptualizations reflect a continuity with earlier speculations on animals' capacities for inference[50]. Trajectory events offer a new experimental model for the study of these varied functions.

METHODS SUMMARY

A microdrive array containing 40 independently adjustable, gold-plated tetrodes aimed at area CA1 of dorsal hippocampus (20 tetrodes per hemisphere; 4.00 mm posterior and 2.85 mm lateral to bregma) was implanted in four rat subjects. Final tetrode placement and unit recording were as previously described[22].

Position information was binned into 2-cm bins. Tuning curves were calculated as the smoothed histogram of firing activity normalized by the time spent per bin. Population events were defined as peaks in a smoothed spike density histogram greater than the mean + 3 standard deviations, bounded by crossings of the mean.

Probability-based decoding of position information from spike trains was performed as previously described[23], using a time window of 20 ms. Each candidate event was truncated to the longest sequence of time frames in which the peak posterior probability was less than 20 cm from that of the previous frame. Events with fewer than 10 steps in the final sequence or a start-to-end distance less than 40 cm were eliminated from further analysis.

Full Methods and any associated references are available in the online version of the paper.

Received 28 September 2012; accepted 21 March 2013.
Published online 17 April 2013.

1. O'Keefe, J. & Nadel, L. *The Hippocampus As A Cognitive Map*. (Clarendon, 1978).
2. Morris, R. G., Garrud, P., Rawlins, J. N. & O'Keefe, J. Place navigation impaired in rats with hippocampal lesions. *Nature* **297,** 681–683 (1982).
3. Scoville, W. B. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* **20,** 11–21 (1957).
4. Olton, D. S. & Samuelson, R. J. Remembrance of places past: spatial memory in rats. *J. Exp. Psychol. Anim. Behav. Process.* **2,** 97–116 (1976).
5. Levy, W. B. A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus* **6,** 579–590 (1996).
6. Redish, A. D. & Touretzky, D. S. The role of the hippocampus in solving the Morris water maze. *Neural Comput.* **10,** 73–111 (1998).
7. Koene, R. A., Gorchetchnikov, A., Cannon, R. C. & Hasselmo, M. E. Modeling goal-directed spatial navigation in the rat based on physiological data from the hippocampal formation. *Neural Netw.* **16,** 577–584 (2003).
8. Foster, D. J. & Knierim, J. J. Sequence learning and the role of the hippocampus in rodent navigation. *Curr. Opin. Neurobiol.* **22,** 294–300 (2012).
9. Hok, V. *et al.* Goal-related activity in hippocampal place cells. *J. Neurosci.* **27,** 472–482 (2007).
10. Wood, E. R., Dudchenko, P. A., Robitsek, R. J. & Eichenbaum, H. Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron* **27,** 623–633 (2000).
11. Ferbinteanu, J. & Shapiro, M. L. Prospective and retrospective memory coding in the hippocampus. *Neuron* **40,** 1227–1239 (2003).
12. Wilson, M. A. & McNaughton, B. L. Dynamics of the hippocampal ensemble code for space. *Science* **261,** 1055–1058 (1993).
13. O'Keefe, J. & Recce, M. L. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3,** 317–330 (1993).
14. Muller, R. U. & Kubie, J. L. The firing of hippocampal place cells predicts the future position of freely moving rats. *J. Neurosci.* **9,** 4101–4110 (1989).
15. Skaggs, W. E., McNaughton, B. L., Wilson, M. A. & Barnes, C. A. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6,** 149–172 (1996).

16. Jensen, O. & Lisman, J. E. Hippocampal CA3 region predicts memory sequences: accounting for the phase precession of place cells. *Learn. Mem.* **3,** 279–287 (1996).
17. Foster, D. J. & Wilson, M. A. Hippocampal theta sequences. *Hippocampus* **17,** 1093–1099 (2007).
18. Johnson, A. & Redish, A. D. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27,** 12176–12189 (2007).
19. Johnson, A., van der Meer, M. A. & Redish, A. D. Integrating hippocampus and striatum in decision-making. *Curr. Opin. Neurobiol.* **17,** 692–697 (2007).
20. Louie, K. & Wilson, M. A. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29,** 145–156 (2001).
21. Lee, A. K. & Wilson, M. A. Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* **36,** 1183–1194 (2002).
22. Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440,** 680–683 (2006).
23. Davidson, T. J., Kloosterman, F. & Wilson, M. A. Hippocampal replay of extended experience. *Neuron* **63,** 497–507 (2009).
24. Diba, K. & Buzsaki, G. Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neurosci.* **10,** 1241–1242 (2007).
25. Karlsson, M. P. & Frank, L. M. Awake replay of remote experiences in the hippocampus. *Nature Neurosci.* **12,** 913–918 (2009).
26. Gupta, A. S., van der Meer, M. A., Touretzky, D. S. & Redish, A. D. Hippocampal replay is not a simple function of experience. *Neuron* **65,** 695–705 (2010).
27. Steele, R. J. & Morris, R. G. Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. *Hippocampus* **9,** 118–136 (1999).
28. Olypher, A. V., Lansky, P. & Fenton, A. A. Properties of the extra-positional signal in hippocampal place cell discharge derived from the overdispersion in location-specific firing. *Neuroscience* **111,** 553–566 (2002).
29. Kentros, C. G., Agnihotri, N. T., Streater, S., Hawkins, R. D. & Kandel, E. R. Increased attention to spatial context increases both place field stability and spatial memory. *Neuron* **42,** 283–295 (2004).
30. Eichenbaum, H., Otto, T. & Cohen, N. J. The hippocampus–what does it do? *Behav. Neural Biol.* **57,** 2–36 (1992).
31. Foster, D. J., Morris, R. G. & Dayan, P. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* **10,** 1–16 (2000).
32. Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A. D. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* **131,** 1–11 (2005).
33. Csicsvari, J., O'Neill, J., Allen, K. & Senior, T. Place-selective firing contributes to the reverse-order reactivation of CA1 pyramidal cells during sharp waves in open-field exploration. *Eur. J. Neurosci.* **26,** 704–716 (2007).
34. Girardeau, G., Benchenane, K., Wiener, S. I., Buzsaki, G. & Zugaro, M. B. Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neurosci.* **12,** 1222–1223 (2009).
35. Ego-Stengel, V. & Wilson, M. A. Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus* **20,** 1–10 (2010).
36. Jadhav, S. P., Kemere, C., German, P. W. & Frank, L. M. Awake hippocampal sharp-wave ripples support spatial memory. *Science* **336,** 1454–1458 (2012).
37. Jackson, J. & Redish, A. D. Network dynamics of hippocampal cell-assemblies resemble multiple spatial maps within single tasks. *Hippocampus* **17,** 1209–1229 (2007).
38. Buzsáki, G. Two-stage model of memory trace formation: a role for "noisy" brain states. *Neuroscience* **31,** 551–570 (1989).
39. Mehta, M. R., Lee, A. K. & Wilson, M. A. Role of experience and oscillations in transforming a rate code into a temporal code. *Nature* **417,** 741–746 (2002).
40. Gerstner, W. & Abbott, L. F. Learning navigational maps through potentiation and modulation of hippocampal place cells. *J. Comput. Neurosci.* **4,** 79–94 (1997).
41. Poucet, B., Thinusblanc, C. & Chapuis, N. Route planning in cats, in relation to the visibility of the goal. *Anim. Behav.* **31,** 594–599 (1983).
42. Foster, D. & Dayan, P. Structure in the space of value functions. *Mach. Learn.* **49,** 325–346 (2002).
43. Sutton, R. S. in *Neural Networks for Control* (eds Miller, T., Sutton, R. S. & Werbos, P.) Ch. 8 (MIT Press, 1990).
44. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16,** 1936–1947 (1996).
45. Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L. & Pennartz, C. M. Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* **7,** e1000173 (2009).
46. van der Meer, M. A., Johnson, A., Schmitzer-Torbert, N. C. & Redish, A. D. Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* **67,** 25–32 (2010).
47. Tulving, E. Episodic memory: from mind to brain. *Annu. Rev. Psychol.* **53,** 1–25 (2002).
48. Hassabis, D., Kumaran, D., Vann, S. D. & Maguire, E. A. Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl Acad. Sci. USA* **104,** 1726–1731 (2007).
49. Buckner, R. L. The role of the hippocampus in prediction and imagination. *Annu. Rev. Psychol.* **61,** 27–48 (2010).
50. Tolman, E. C. *Purposive Behavior in Animals and Men* (The Century Co., 1932).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Alfred P. Sloan Foundation, The Brain and Behavior Research Foundation (NARSAD Young Investigator Grant) and the National Institutes of Health grant MH085823.

Author Contributions B.E.P. and D.J.F. designed the experiment and analyses, B.E.P. collected the data, B.E.P. and D.J.F. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.J.F. (david.foster@jhu.edu).

METHODS

Behaviour and data acquisition. All procedures were approved by the Johns Hopkins University Animal Care and Use Committee and followed US National Institutes of Health animal use guidelines. Behavioural training and in-session recording took place from late afternoon to early evening (rats were housed on a standard, non-inverted, 12-h light cycle).

Adult male Long-Evans rats (10–20 weeks old, 450–550 g) were handled daily and food-restricted to 85–90% of their free-feeding weight and then trained to traverse a 1.8-m linear track to receive a liquid chocolate-flavoured reward (200 μl, Carnation) at either end. Rats were trained for the briefer of 20 min or 20 complete laps once per day for at least 10 consecutive days. Linear track training occurred in a room separate and visually distinct from the recording room.

After a rat achieved criterion performance on the linear track (three consecutive days with 20 laps in under 20 min), training on the open field was initiated in a 2 m × 2 m black arena with 30-cm-high walls and 36 identical, evenly spaced, 1.5-cm-diameter, 3-mm-deep conical reward delivery wells embedded into the floor such that the rim of each well was level with the floor (Fig. 1a). Each well was attached to a tubing system that ran beneath the environment, which allowed any well to be independently and soundlessly filled or emptied by the experimenter via a hand-held syringe. During the filling of a well, no obvious visible or audible cue was available to the rat signifying that a well had been filled. When active, wells were filled with 300 μl of chocolate milk. Open-field training took place in the recording room with all room and environmental cues positioned as they would be during the eventual in-session recording.

Open-field training proceeded in four stages. First, each rat underwent one 30-min-long session per day for 2 days in which every available well was filled (and immediately refilled following consumption) and food crumbs were scattered throughout the arena to encourage initial exploration. This was the only stage of training in which non-liquid food was present in the arena. In the second stage of training (3 days), each 30-min-long session began with four filled wells, one per quadrant of the arena. When the reward in one quadrant was consumed, another random well in that quadrant was filled, but only after the rat had left the quadrant and consumed reward from another quadrant. In the third stage (3 days), the final experimental procedure (see below) was begun except that on the interleaved Random trials, two randomly selected wells were filled to make the task easier to complete. When one Random well was discovered and consumed, the second was immediately emptied and the Home well was filled. Finally, on the fourth stage, the rats were trained on the final experimental protocol for the lesser of thirty minutes or for 30 trials until they reached criterion performance (30 trials in less than 30 min for three consecutive days). Every session began by placing the rat in one corner of the arena and then allowing free exploration.

In the final experimental protocol, the Home well was initially filled and was the only filled well in the arena at the start of the session. Once the rat discovered and consumed the Home well reward, a randomly selected well was filled. Only after the rat discovered and consumed the Random well reward was the Home well again filled. A trial consisted of the rat leaving the Home location, discovering and consuming the reward at a Random well and then returning to the Home location and consuming the reward there. At no point in the training were the rats provided with any cue informing them when the Home or a Random well was filled (filling occurred during or immediately after consumption at the prior well). Instead, the rats learned to return to the Home well location without cue after consuming the reward at a filled Random well and to begin searching for a Random well immediately after consuming the reward at Home. The Home well location changed every session, but was constant throughout the session. The location of the Home well on the recording days had never previously been experienced by the rats as a Home well location, although they had sporadically received reward in those locations as Random wells in previous sessions.

After a rat achieved criterion performance on the task, it was surgically implanted with a microdrive array (25–30 g) containing 40 independently adjustable, gold-plated tetrodes aimed at area CA1 of dorsal hippocampus (20 tetrodes in each hemisphere; 4.00 mm posterior and 2.85 mm lateral to bregma). Following surgical implantation, tetrodes were slowly lowered into the CA1 pyramidal layer over the course of 7–10 days. Final tetrode placement and unit recording were as previously described[22]. Each tetrode consisted of a twisted bundle of four 17.8 μm platinum/10% iridium wires (Neuralynx), and each wire was electroplated with gold to an impedance of <150 MΩ before surgery. A bone screw firmly attached to the skull served as ground. During the first 4 or 5 days following implantation, the rat was not re-exposed to the experimental arena. After this recovery time, while tetrodes were still being advanced to the hippocampus, the rat was trained once per day on the final experimental protocol for the lesser of 30 min or 30 trials to familiarize it with navigating the arena with the microdrive and attached wires.

All data were collected using a Neuralynx data acquisition system and an overhead video system that recorded continuously at 60 Hz. The rat's position and head direction were determined via two distinctly coloured, head-mounted LEDs. Analogue neural signals were digitized at 32,556 Hz. Spike threshold crossings (50 μV) were recorded at 32,556 Hz. Continuous local field potential data were digitally filtered between 0.1 and 500 Hz and recorded at 3,255.6 Hz. The beginning and end of reward consumption were manually determined from the captured video data.

Cluster analysis. Individual units were identified by manual clustering based on spike waveform peak amplitudes using custom software (xclust2, M. A. Wilson). Only well-isolated units were included in the analysis. Modified L_{ratio} values[32] were calculated for each cluster to confirm cluster quality using the peak amplitude of each waveform as the feature set. Briefly, the L_{ratio} value of cluster C is

$$L_{\text{ratio}}=\left(\sum_{i\notin C}\left(1-\mathrm{CDF}_{\chi^2_{df}}\left(D^2_{i,C}\right)\right)\right)/n_s$$

where n_s is the total number of spikes recorded on the tetrode throughout the experiment, $i\notin C$ is the set of spikes which are not members of cluster C, $D^2_{i,C}$ is the Mahalanobis distance of spike i from cluster C, and $\mathrm{CDF}_{\chi^2_{df}}$ is the cumulative distribution function of the χ^2 distribution with $df=4$. We modified the original equation for L_{ratio} to allow for comparison between tetrodes with different numbers of spikes and between experiments of varying time spans. As the original equation is a sum, even well-isolated clusters will necessarily have larger L_{ratio} values for particularly long experimental sessions or if they occur on tetrodes with large numbers of spikes. Thus, we normalized the sum by the total number of spikes recorded on the tetrode.

Clustered units that may correspond to putative inhibitory neurons were excluded on the basis of spike width and mean firing rate. To ensure accurate decoding of hippocampal events, only rats in which we obtained at least 100 simultaneously recorded place units were used for subsequent analysis.

Decoding spatial location. Position was binned (2 cm) and position tuning curves (place fields) were calculated as the smoothed (Gaussian kernel, standard deviation of 4 cm) histogram of firing activity normalized by the time spent per bin. Only periods of time when the rat was moving faster than 5 cm s^{-1} were used to determine place fields. Units were considered to have a place field if the unit was classified as excitatory and the peak of the tuning curve was >1 Hz.

A memoryless probability-based decoding algorithm[23] was used to estimate the rat's position throughout the experiment based on the unit position tuning curves and the spike trains. Briefly, the probability of the animal's position (pos) across M total position bins given a time window (τ) containing neural spiking (spikes) is

$$\Pr(\text{pos}|\text{spikes})=\cup/\sum_{j=1}^{M}\cup$$

where

$$\cup=\left(\prod_{i=1}^{N}f_i(\text{pos})^{n_i}\right)e^{-\tau\sum_{i=1}^{N}f_i(\text{pos})}$$

and $f_i(\text{pos})$ is the position tuning curve of the i-th unit, assuming independent rates and Poisson firing statistics for all N units and a uniform prior over position. A time window of 250 ms was used to estimate the rat's position on a behavioural timescale. A time window of 20 ms was used to estimate position during candidate population events.

Sequential event analysis. A histogram (1-ms bins) of all clustered units for times when the rat's velocity was less than 5 cm s^{-1} was smoothed (Gaussian kernel, standard deviation of 10 ms). Population events were defined as peaks in the smoothed histogram greater than the mean + 3 standard deviations. Start and end boundaries for each population event were defined as the points where the smoothed histogram crossed the mean. To prevent estimation artefacts, the time window boundaries for each candidate event were adjusted inward (if necessary) to ensure that the first and last estimation bins contained a minimum of 2 spikes. Candidate events in which fewer than 10% of the clustered units participated or with boundaries less than 50 ms or greater than 2,000 ms apart were excluded from analysis.

For each candidate event, the rat's position was estimated using the probability-based decoding algorithm described above with a 20-ms time window, advanced in 5-ms increments throughout the putative event. Following position estimation, each candidate replay event was truncated to the longest sequence of time frames with peak posterior probability less than 20 cm from that of the previous frame. Candidate events with fewer than 10 steps in the final sequence or a start-to-end distance less than 40 cm were eliminated from future analysis. The remaining candidate events were categorized as 'trajectory events'.

For trajectory event quantification, the posterior probabilities for every time frame of each trajectory event were summed across time. For comparison between away-events and home-events, these sums were normalized for the number of time-frames in each event. For all analyses requiring per-well quantification, the arena was subdivided by drawing an imaginary line equidistant between each well, resulting in 36 regions, each encompassing an approximately 33×33 cm area (Supplementary Fig. 4). Quantification for all event trajectory analysis in which the rat's location was not specifically examined did not include the area within 15 cm of the rat's physical location at the time of the event to avoid initiation bias.

For all trajectory events, a Monte-Carlo *P*-value was calculated using two shuffle methods: randomly shuffling cell identity and randomly shuffling each cell's place field in both the x and y dimensions. The *P*-value was calculated as $(n+1)/(r+1)$, where n is the number of shuffles that met the criteria to be classified as a trajectory event and r is the total number of shuffles. 5,000 shuffles were used for both methods. All candidate events that met our criteria to be classified as trajectory events had a *P*-value less than 0.02 for both shuffle methods.

To quantify the precise spatial correlation between trajectory events and the rat's future/past path, each trajectory event was transformed into a vector of the peak posterior probabilities for each time frame of the event. Using the rat's physical location at the time of the event as the centre, concentric rings were drawn around the rat with radial increments of 2 cm, starting with a radius of 15 cm. For each ring, the first crossing for the event vector and the rat's future or past path were determined and the angular displacement (the minor arc along the ring's circumference, normalized by the ring's radius) was calculated between these points. This value was compared to that obtained from 2,000 randomly selected events (chosen from across all sessions) which were spatially relocated so that the rat's physical location at the time of the random event matched the rat's physical location at the time of the trajectory event to generate a Monte-Carlo *P*-value.

Local field potential analysis. For each tetrode, one representative electrode was selected and the local field potential signal was analysed. To examine SWRs, the local field potential was band-pass filtered between 150 and 250 Hz, and the absolute value of the Hilbert transform of this filtered signal was then smoothed (Gaussian kernel, s.d. = 12.5 ms). This processed signal was averaged across all tetrodes and ripple events were identified as local peaks with an amplitude greater than 3 s.d. above the mean, using only periods when the rat's velocity was less than 5 cm s^{-1}. The start and end boundaries for each event were defined as the point when the signal crossed the mean. For theta-band power analysis, the raw local field potential trace was band-pass filtered between 4 and 12 Hz and the absolute value of the Hilbert transform of the filtered signal was calculated. The z-score theta power for each electrode was determined for every time point of the 60 Hz position data and for 100–200 ms before and after each identified trajectory event. For power spectral density analysis, 100 ms non-overlapping temporal bins were used to compute the spectrograms. A z-score was calculated for each frequency band across the entire behavioural session. The SWR or trajectory event triggered spectrograms use the peak of the ripple power or the peak of the spike density, respectively, as time zero.

doi:10.1038/nature12104

Structures of the human and *Drosophila* 80S ribosome

Andreas M. Anger[1]*, Jean-Paul Armache[1]*, Otto Berninghausen[1], Michael Habeck[2,3], Marion Subklewe[4], Daniel N. Wilson[1] & Roland Beckmann[1]

Protein synthesis in all cells is carried out by macromolecular machines called ribosomes. Although the structures of prokaryotic, yeast and protist ribosomes have been determined, the more complex molecular architecture of metazoan 80S ribosomes has so far remained elusive. Here we present structures of *Drosophila melanogaster* and *Homo sapiens* 80S ribosomes in complex with the translation factor eEF2, E-site transfer RNA and Stm1-like proteins, based on high-resolution cryo-electron-microscopy density maps. These structures not only illustrate the co-evolution of metazoan-specific ribosomal RNA with ribosomal proteins but also reveal the presence of two additional structural layers in metazoan ribosomes, a well-ordered inner layer covered by a flexible RNA outer layer. The human and *Drosophila* ribosome structures will provide the basis for more detailed structural, biochemical and genetic experiments.

Crystal structures of prokaryotic ribosomal particles have provided insights into protein biosynthesis at both a structural and a functional level[1]. In contrast to their bacterial counterparts, eukaryotic ribosomes are much larger and more complex; they contain approximately 2,650 nucleotides of additional rRNA in *H. sapiens* in the form of so-called expansion segments and 26 additional ribosomal proteins as well as 2,452 amino acids of ribosomal protein extensions[2–4]. Cryo-electron microscopy (cryo-EM)[5–7] and crystal structures[8–10] have elucidated the architecture of yeast, protist and plant ribosomes. In contrast, the limited resolution (9 to 20 Å) of cryo-EM structures of mammalian 80S ribosomes[11–14] has so far prohibited the generation of complete molecular models for these metazoans.

Here we present single-particle cryo-EM structures of monomeric 80S ribosomes isolated from *D. melanogaster* embryonic extracts and human peripheral blood mononuclear cells (Supplementary Fig. 1). *In silico* sorting was used to generate homogeneous data sets with additional density corresponding to eukaryotic elongation factor 2 (eEF2), in agreement with mass spectrometry analysis (Supplementary Tables 1 and 2). The eEF2-containing particles seemed to be stabilized in a rotated conformation, allowing the reconstructions of each sub-data set to reach an average resolution of 5.4 to 6.0 Å (Supplementary Fig. 2). Notably, local resolution of the human 80S ribosome ranged from above 9 Å on the flexible periphery to better than 4.8 Å for large parts of the map (Fig. 1a). This is in agreement with the distinct structural details observed throughout the map: the pitch of α-helices is visible and strand-separation is recognizable for many β-sheets of ribosomal proteins (Fig. 1b). Density is also visible for a number of bulky side chains (Fig. 1b). In terms of rRNA, the phosphate–ribose backbone is well resolved and bulged-out bases are clearly represented (Fig. 1c). Moreover, the quality of the cryo-EM map enabled us to distinguish between human rRNA sequence variations (Supplementary Fig. 3). Our electron-density maps, coupled with secondary structure predictions for the rRNA expansion segments and the available yeast and *Tetrahymena* crystal structures[8–10], enabled us to build complete molecular models for both the *Drosophila* and human 80S ribosome (Fig. 1d, e and Supplementary Tables 3–8).

Ribosomal protein extensions

With the exception of yeast, which lacks L28e, eukaryotic cytoplasmic 80S ribosomes contain the same set of 80 core ribosomal proteins (Fig. 2a, b, Supplementary Fig. 4 and Supplementary Tables 3–6). Compared to yeast and protists, there has been a modest increase in protein mass in metazoan ribosomes, specifically by a total of 1,094 amino acids (approximately 8%) and 796 amino acids (approximately 6%) in the *Drosophila* and human 80S ribosomes, respectively. On the 40S subunit, the protein mass increase of *Drosophila* (210 amino acids, approximately 4%) and human (147 amino acids, 3%) relative to yeast is small, and mostly disordered in the cryo-EM maps. Notable exceptions include the carboxy-terminal extension (CTE) of S26e, which reaches into the messenger RNA exit channel (Supplementary Fig. 5), and part of the CTE of S6e that bridges the right and left feet of the 40S subunit (Fig. 2b and Supplementary Fig. 5). Phosphorylation of the CTE of S6e by S6 kinase (S6K) is important for translation regulation, as well as glucose homeostasis and regulation of cell size in metazoans[15]. The S6K consensus recognition motif (RXRXXS), which was disordered in recent X-ray structures of the yeast 80S ribosome and *Tetrahymena* 40S subunit[8,10], adopts an α-helical conformation in the human 80S ribosome (Supplementary Fig. 5). The most dramatic increases in ribosomal protein extensions are seen on the 60S subunits for ribosomal proteins L4, L6e, L14e and L29e, as well as for *Drosophila* L22e and L23. Collectively, these account for 52% (460 amino acids) and 58% (375 amino acids) of the total protein mass gain in the *Drosophila* and human 60S subunit, respectively. Notably, the approximately 180- and 140-amino-acid extensions of L22e and L23, respectively, double the size of these *D. melanogaster* ribosomal proteins, compared to other non-insect species such as yeast and human (Supplementary Fig. 6). Structures of yeast and *Tetrahymena* ribosomes revealed a highly complex network of RNA–protein interactions between the eukaryote-specific ribosomal protein extensions and the rRNA expansion segments[2–4,7–10]. The dimensions of this RNA–protein layer has developed further in metazoan ribosomes, which is illustrated by the increasing size and complexity of the interaction between expansion segment 7L

[1]Gene Center, Department of Biochemistry and Center for integrated Protein Science Munich (CiPSM), Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 Munich, Germany. [2]Department of Empirical Inference, Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tübingen, Germany. [3]Department of Protein Evolution, Max Planck Institute for Developmental Biology, Spemannstrasse 38, 72076 Tübingen, Germany. [4]Department of Internal Medicine III, Klinikum der Universität München and Clinical Cooperation Group Immunotherapy at the Helmholtz Institute Munich, Marchioninistrasse 15, 81377 Munich, Germany.
*These authors contributed equally to this work.

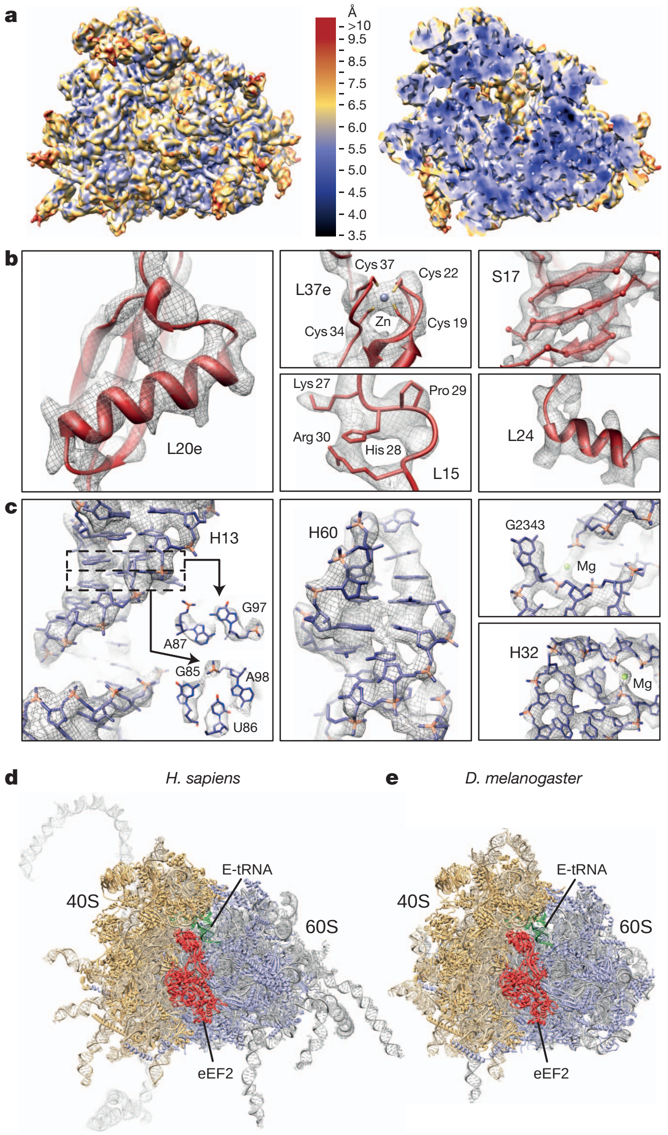


Figure 1 | Structures of the human and *Drosophila* 80S ribosomes. **a**, Surface and cross section of the human 80S ribosome electron density map (filtered at 6 Å for clarity), coloured according to the local resolution. **b**, **c**, Selected views of the *H. sapiens* 80S map (grey mesh) with (**b**) protein and (**c**) rRNA. RNA backbone phosphates are highlighted in orange. **d**, **e**, Complete models of the human and *Drosophila* 80S ribosomes with ribosomal proteins and rRNA of the 40S and 60S subunits shown in orange and blue, respectively. Flexible human ES27L (light grey) is shown in an arbitrary position.

(ES7L) with the NTE of L6e (Supplementary Fig. 7). Interestingly, the extensions of human L4, L14e and L29e and *Drosophila* L22e and L23, show similarity to the flexible C-terminal regions of the linker histone H1 in that they are highly basic and enriched in alanine, lysine and proline residues[16] (Supplementary Fig. 8). The histone H1 tails have been proposed to form α-helical conformations punctuated by proline breaks, which track one groove of the linker DNA (reviewed previously[17]). In the *Drosophila* and human 80S ribosome, it seems that these histone H1-like ribosomal protein parts are directed towards adjacent expansion segments. However, owing to the flexibility of the expansion segments, it was not possible to model the associated extensions (Supplementary Fig. 8).

Ribosomal proteins, eEF2 and Stm1-like factors

As in yeast and *Tetrahymena* extensions of *Drosophila* and human ribosomal proteins S13, S19, S25e, S30e and S31e extend into the functional centre of the 40S subunit. There, the amino-terminal extension (NTE) of ribosomal proteins S30e and S31e establish interactions with eEF2 (Supplementary Fig. 9). This was not observed in the lower resolution yeast eEF2–80S complexes[5,12,18]. Moreover, although the overall conformation and contacts of *Drosophila* and human eEF2 on the ribosome are very similar to those observed for yeast[5,12,18] (Fig. 2c, d, Supplementary Fig. 10, and Supplementary Tables 9 and 10), at higher resolution we could also model interactions between the N-terminal domain of L11, domain II of the L10 stalk protein and the G′ domain of human and *Drosophila* eEF2 (Fig. 2e). Additional density is present in the human eEF2 for the mammal-specific insertion within the G′ domain. This additional density is absent in *Drosophila* eEF2 (Fig. 2f). At lower thresholds, extra density is observed adjacent to this region. This may represent the C terminus of the 60S acidic ribosomal P1 and P2 stalk proteins, reminiscent of the interaction between the bacterial L7 and L12 stalk proteins and the G′ domain of EF-G[19,20]. In contrast to bacterial EF-G, archaeal EF2s and eEF2s are post-translationally modified by conversion of a conserved histidine (His 699, His 701 and His 715 for yeast, *Drosophila*, and human eEF2, respectively) to diphthamide. Deletion of the modification enzymes in mice leads to embryonic lethality or severe developmental defects[21]. Moreover, diphthamide is adenosine di-phosphate (ADP)-ribosylated by the diphtheria toxin, which inactivates eEF2 and inhibits protein synthesis[21]. In the human 80S–eEF2 structure, we observe density for the diphthamide residue contacting the backbone of H44 in the vicinity of A1825 (A1493 in *Escherichia coli* numbering) (Fig. 2g). In bacteria, A1492 and A1493 are involved in recognition of the mRNA–tRNA duplex during decoding[22,23], thus contact of diphthamide with this region is consistent with its proposed role to disrupt the interaction between the decoding centre and mRNA–tRNA duplex during translocation[18]. Notably, we also observe an alternative conformation of diphthamide directed towards density located within the path of the mRNA, which we have assigned to the serpine 1 mRNA-binding protein 1 (SERBP1; also known as plasminogen activator inhibitor 1 RNA-binding protein) based on mass-spectrometry analysis (Fig. 2c, g and Supplementary Table 1). SERBP1 was identified, together with ribosomal proteins and eIF3, to interact with the hepatitis C virus internal ribosomal entry site (IRES)[24], which engages the small ribosomal subunit during initiation[25]. Moreover, SERBP1 is homologous to the translation repressor Stm1 (ref. 26), which is present in the crystal structure of the yeast 80S ribosome purified under conditions of nutrient deprivation[10]. We observe that, like Stm1, SERBP1 has an extended structure passing through the P- and A-tRNA binding sites (Fig. 2c, d); it then follows the mRNA channel to the solvent side, where it interacts with ribosomal proteins S5, S10e, S12e and S30e located on the head of the 40S subunit (Fig. 2c and Supplementary Figs 10 and 11, and Supplementary Table 11). Examination of the *Drosophila* 80S ribosome also revealed a density within these regions, which was identified by mass spectrometry to be VIG2 (Supplementary Tables 2 and 12), a protein orthologous to SERBP1 (Supplementary Figs 10 and 11). The identification of SERBP1 and VIG2 on metazoan 80S ribosomes indicates a novel role, analogous to Stm1 in yeast, for these proteins in the regulation of translation in *Drosophila* and humans.

Ribosomal RNA expansion segments

We were able to localize and build models for all 30 rRNA expansion segments (we use an extended nomenclature based on a previous paper[27], Supplementary Fig. 12 and Supplementary Tables 7 and 8) of the *Drosophila* and human 80S ribosome, 9 expansion segments of the 40S subunit and 21 expansion segments of the 60S subunit (Fig. 3 and Supplementary Figs 13–16). Although human and *Drosophila* contain a similar set of expansion segments as yeast and protists, their expansion segments are generally much longer, exemplified by comparing ES3S, ES7L, ES9L, ES15L, ES27L and ES39L between yeast (approximately 110, 200, 70, 20, 160 and 140 nucleotides) and human (longer by 50, 670, 40, 170, 550 and 100 nucleotides) (Supplementary Tables 7 and 8). In addition, metazoans contain ES30L and ES43L (Fig. 3), which are lacking in yeast and *Tetrahymena*. Although the

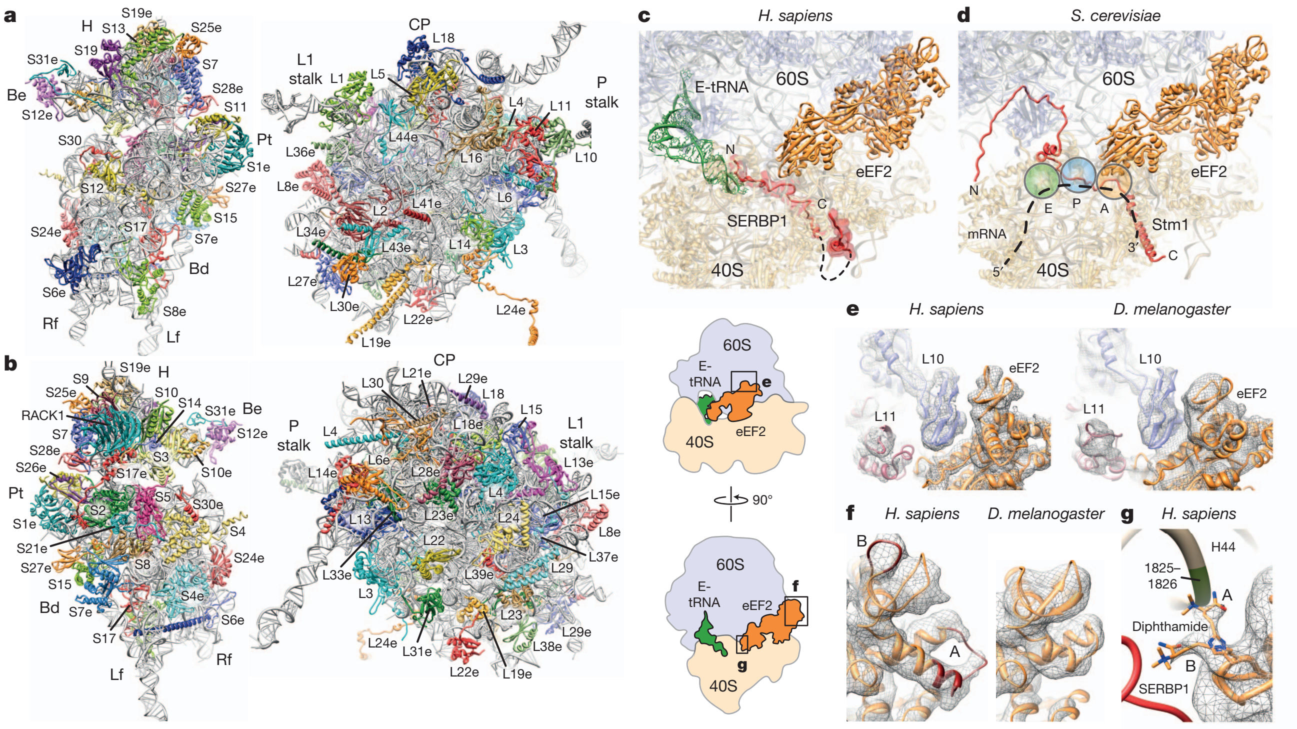


Figure 2 | Protein architecture of the human 80S ribosome and associated factors. **a**, **b**, Interface (**a**) and solvent (**b**) view of the human 40S (left) and 60S (right) ribosome subunits, with rRNA shown in grey and ribosomal proteins coloured individually. Be, beak; Bd, body; CP, central protuberance; H, head; Lf, left foot; Pt, platform; Rf, right foot. **c**, Relative position of eEF2 (orange), E-site tRNA (green) and SERBP1 (red) on the *H. sapiens* 80S ribosome **d**, eEF2[18] and Stm1 (red) in *S. cerevisiae*[10]. Positions of aminoacyl (A), peptidyl (P) and exit (E) tRNA-binding sites are indicated. **e**, Interaction of L11 and L10 with eEF2. **f**, G′ domains of eEF2 with human insertions (A and B, red). **g**, Alternative conformations of the diphthamide–His 715 of eEF2, contacting nucleotides 1825 and 1826 in H44 or SERBP1. The insets show the locations of the areas enlarged in parts **e**, **f** and **g**.

distal ends of several large human rRNA insertions (for example, ES3S, ES6S, ES7L, ES15L, ES27L, ES30L and ES39L) could only be partially resolved in the cryo-EM reconstructions (Supplementary Fig. 17), the flexible tentacle-like nature of these expansion segments was observable within individual electron-microscopy images (Supplementary Fig. 18)[28]. The extreme base composition of expansion segments, being AU-rich in *Drosophila* (32% GC) and GC-rich in human (80% GC) (Supplementary Tables 7, 8 and 13), has prevented secondary structure prediction for approximately 720 (33%) and 1,800 (57%) nucleotides of several expansion segments, respectively[29,30] (Supplementary Figs 19 and 20). However, using iterative model building and focused secondary structure predictions, we conclude that the distal ends of the flexible expansion segments adopt simple, unbranched A-form helices, enabling us to present complete molecular models

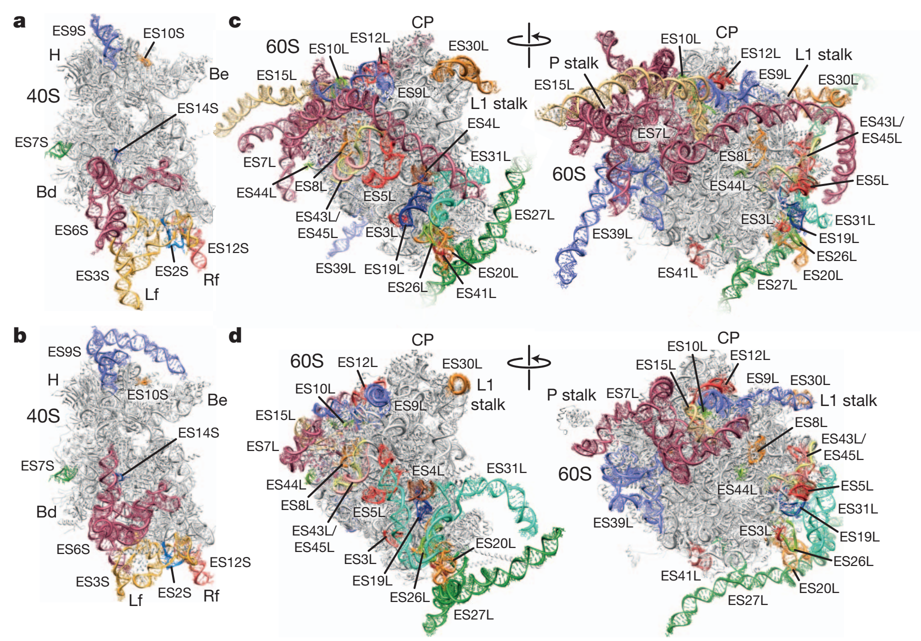


Figure 3 | Metazoan rRNA expansion segments. **a**, **b**, Molecular models of the 40S subunits of (**a**) *H. sapiens* and (**b**) *D. melanogaster* with expansion segments. **c**, **d**, Molecular models of the 60S ribosome subunits of (**c**) *H. sapiens* and (**d**) *D. melanogaster* showing expansion segments.

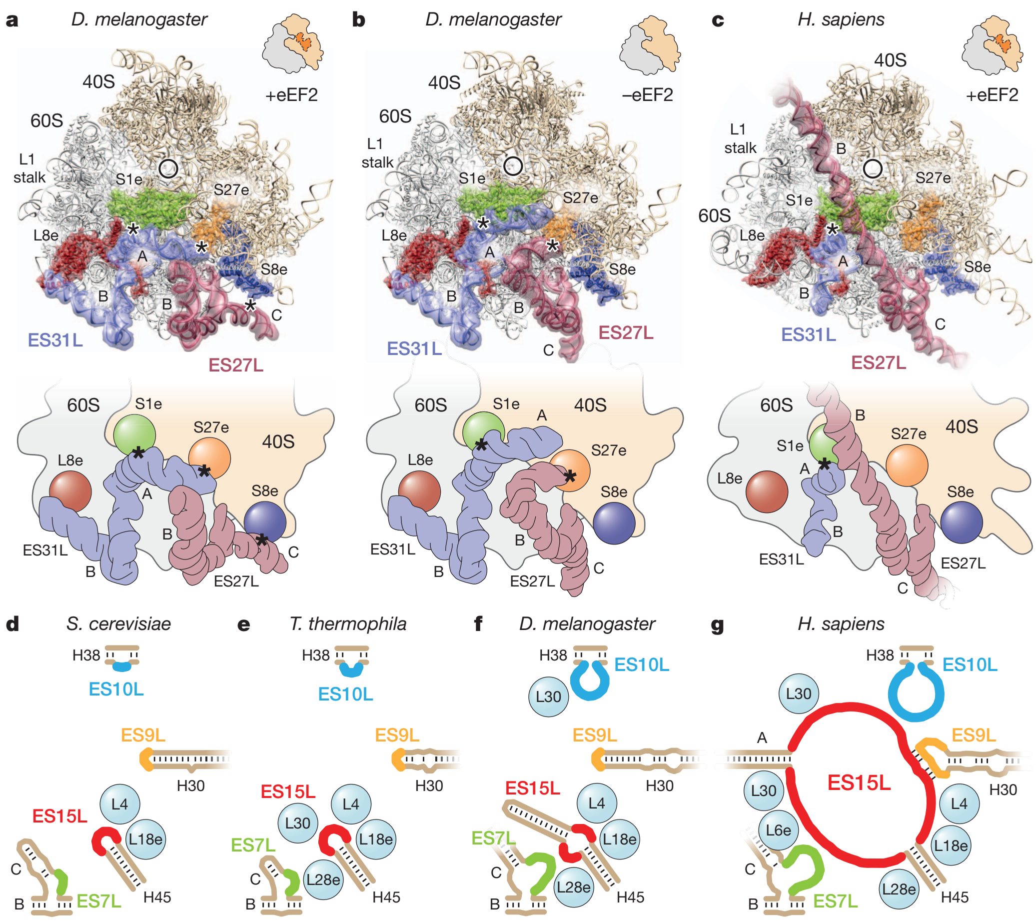


Figure 4 | Dynamic behaviour and co-evolution of expansion segments. **a–c**, Comparison of the ES27L and ES31L behaviour in the eEF2-bound (rotated) (**a**, **c**) and empty (-eEF2, unrotated) (**b**) form of the *Drosophila* and *H. sapiens* ribosome. Bridges with ribosomal proteins are highlighted with asterisks, the mRNA exit site is indicated with a circle. **d–g**, Schematic view comparing the interactions within the expansion-segment cluster formed by ES7L, ES9L, ES10L and ES15L between *S. cerevisiae* (**d**)[6,7,10], *T. thermophila* (**e**)[9], *D. melanogaster* (**f**) and *H. sapiens* (**g**). Non-helical elements of expansion segments are highlighted, and helices are labelled A to C.

(Fig. 3) and refined secondary structure diagrams for the entire human and *Drosophila* small and large subunit rRNAs (Supplementary Figs 13–16).

On the human and *Drosophila* 40S subunits, the expansion segments cluster at the bottom of the back of the particle, where ES3S and ES6S interact tightly (Fig. 3a, b). The terminal loop of helix E of ES6S (ES6S-E) forms continuous base pairs with an internal loop of ES3S-B (Supplementary Fig. 21), as reported previously for yeast, wheat germ and *Tetrahymena*[6,8,10,31]. ES3S-B is extended in human compared to *Drosophila*, yeast and *Tetrahymena*, resulting in a longer left foot of the human 40S subunit (Fig. 3a and Supplementary Fig. 21). Conversely, ES9S is elongated in *Drosophila* and forms a 'horn' that interacts with S31e, thereby spanning the *Drosophila* 40S subunit region of the head comprising the binding site of eEF3 in yeast[32] (Fig. 3b and Supplementary Fig. 22). Although ES6S-A and ES6S-B are conserved in length between yeast, protists and metazoans, the conformations of these helices are markedly different between human and *Drosophila* (Fig. 3a, b), and between human, yeast and protists[4,6,8,10] (Supplementary Fig. 21). In addition, *Drosophila* ES6S-B contains a helical insertion resulting in branched ES6S-B1 and ES6S-B2 helices (Supplementary Fig. 21). Notably, the ES3S–ES6S region contributes to the binding site for the eukaryote-specific translation initiation factors eIF3 and eIF4G[33–35], emphasizing that structural variation in this region is likely to reflect functional differences during eukaryotic translation initiation.

Expansion segments of the human and *Drosophila* 60S subunit are mainly positioned on the side and back of the particle, with clusters located adjacent to the L1 and P stalks (Fig. 3c, d). Compared to yeast and protists, the most dramatic increase in mass is formed by ES7L, ES9L, ES10L, ES15L, ES27L and ES39L (Fig. 3c, d). Interestingly, the terminal loop of H30 within ES9L in the human rRNA forms continuous base pairs with an internal part of ES15L (Fig. 3c and Supplementary Fig. 23), analogous to the hybrid helix formed between ES3S–ES6S in the 40S subunit (Fig. 3a, b and Supplementary Fig. 21). The resulting mixed ES9L–ES15L helix seems to anchor the base of the human-specific extension of ES15L tightly to the surface of the particle.

Dynamic behaviour of expansion segments

As in yeast[10], human and *Drosophila* ES31L-A interacts with ribosomal protein S1e on the 40S subunit to form the eukaryote-specific intersubunit bridge eB8 (Fig. 4a–c). *Drosophila* ES31L is approximately 130 nucleotides longer than those of yeast and human (Supplementary Table 8), resulting in a prolonged helix ES31L-B that contacts L8e (Fig. 4a). Furthermore, helix ES31L-A is elongated and establishes a novel intersubunit bridge (which we term eB15, extending the nomenclature of yeast and protist ribosomes[10]) with ribosomal protein S27e near the mRNA exit site on the 40S subunit (Fig. 4a). In *Drosophila*, helix ES27L-C is extended compared to yeast ES27L, resulting in the formation of a second metazoan-specific intersubunit bridge (eB16) through interaction with S8e (Fig. 4a). Although human ES27L is larger than both those of yeast and *Drosophila*, contact to S8e is not observed in the human 80S ribosome because it adopts a conformation extending towards the L1 stalk (ES27L-in) (Fig. 4a–c). In addition to ES27L-in, an ES27L-out conformation that reaches towards

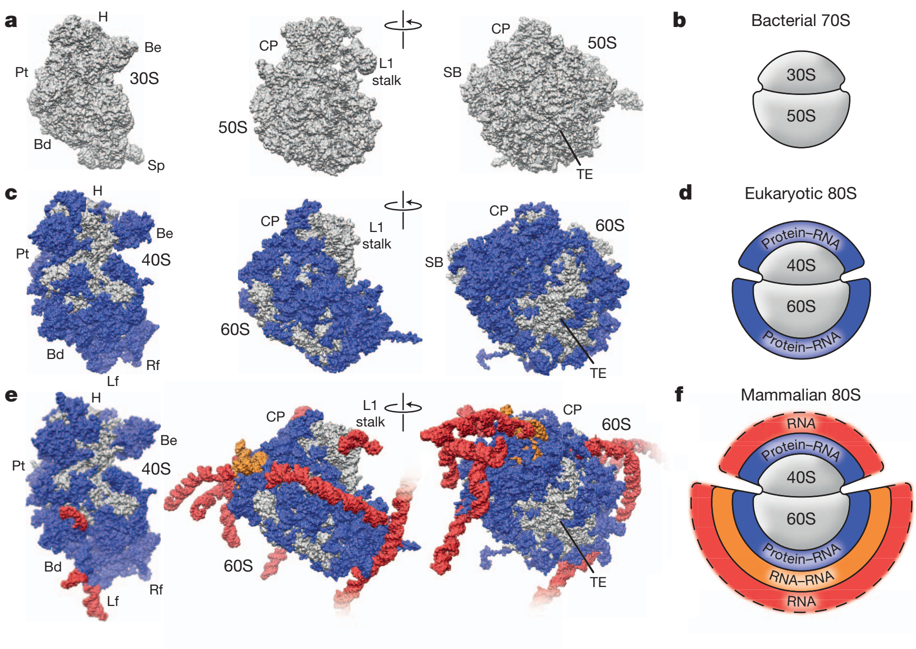


Figure 5 | Layered evolution of the eukaryotic ribosome. a–f, Surface representations (**a**, **c**, **e**) and schematics (**b**, **d**, **f**) of the bacterial *T. thermophilus* 70S ribosome (**a**, **b**)[47], the *S. cerevisiae* 80S ribosome (**c**, **d**)[10] (the eukaryote-specific protein–RNA layer is shown), and the mammalian 80S ribosome from *H. sapiens* (**e**, **f**) (the two additional layers, RNA–RNA and RNA-only, are shown). SB, P-stalk base; Sp, spur; TE, tunnel exit.

the tunnel exit has been observed previously for various yeast 80S ribosome complexes[6,36,37]. However, the *Drosophila* ES27L more closely resembles the intermediate ES27L position observed in wheat germ 80S ribosomes[6]. We therefore analysed the conformation of ES27L in the sub-populations of *Drosophila* and human 80S ribosomes that lacked eEF2 and exhibited non-rotated states. Although the ES27L-in conformation was identical between the eEF2-bound rotated and eEF2-lacking non-rotated human 80S ribosomes, a dynamic interplay of structural rearrangements was observed between *Drosophila* ES27L and ES31L (Fig. 4a, b): in the non-rotated state, we observe an ES27L-in conformation, such that the intersubunit bridge between ES27L-C and S8e is absent (Fig. 4b). In contrast, ES27L-B seems to displace ES31L-A to re-establish an intersubunit bridge with S27e (eB17) (Fig. 4b). Although ES31L-A maintains contact with S1e, the distal end of ES31L-A becomes disordered, presumably owing to the loss of interaction with S27e. The role of the dynamic rearrangements requires further investigation, but it seems that the conformational dynamics of ES27L and ES31L enable communication between two functional important regions of the ribosome, the mRNA exit site on the 40S subunit and the tunnel exit site on the 60S subunit. Indeed, deletion of ES27L in *Tetrahymena* is lethal[38], and the ES27L-out conformation has been observed to interact with a variety of important factors at the tunnel exit site, such as the nuclear export factor Arx1 (refs 39, 40), the ribosome-associated complex[41] and the membrane protein ERJ1 (Erj5p in *S. cerevisiae*)[42].

RNA–RNA interaction

It has been noted that ES31L and ES39L in yeast and *Tetrahymena* ribosomes use extended single-stranded (non-helical) rRNA stretches as platforms for interactions with ribosomal proteins[4,9,10]. In addition to ES31L and ES39L, the same structural principle is even more pronounced in metazoan ribosomes, and non-helical stretches are also observed in ES7L, ES10L and ES15L. Moreover, these structural elements are not only used for protein-RNA interactions but also establish RNA–RNA interactions between the expansion segments (Fig. 4d–g). In yeast and *Tetrahymena*, ES10L represents an asymmetric loop of 3 and 5 nucleotides, respectively, inserted into H38 (Fig. 4d, e). This non-helical insertion of ES10L has increased in *Drosophila* (by 12 nucleotides) and humans (by 22 nucleotides), leading to additional contacts with L30 and ES15L, respectively (Fig. 4f, g). In yeast, the loop of ES15L caps H45 and interacts with L4 and L18e (Fig. 4d), whereas in metazoans the insertion of helix ES15L-A creates an enlarged internal loop (Fig. 4f, g). In the *Drosophila* ribosome, the 9 non-helical nucleotides of this internal loop interact with ribosomal proteins L4, L18e and L28e, and also form contacts with the non-helical insertion of ES7L (Fig. 4f). The internal loop is further enlarged in the human ribosome, leading to new contacts with ribosomal proteins L6e and L30 as well as ES7L, ES9L and ES10L (Fig. 4g). Collectively, it seems that in metazoans, the non-helical insertions form a complex network of RNA–protein and RNA–RNA interactions that contribute to the stabilization of the large expansion segments cluster on the back of the 60S subunit.

Conclusion

The majority of the rRNA and ribosomal proteins that constitute the bacterial 70S ribosome is conserved in eukaryotes, and can therefore be considered to form the core of the 80S ribosome (Fig. 5a, b). Structures of the yeast and *Tetrahymena* ribosomes have revealed that the additional eukaryote-specific ribosomal proteins form a network of interactions with the rRNA expansion segments, resulting in an intertwined RNA–protein layer[6–10] (Fig. 5c, d). In metazoan eukaryotes, this RNA–protein layer has increased in size and complexity owing to the presence of additional ribosomal protein extensions and rRNA expansion-segment insertions (Fig. 5e, f and Supplementary Fig. 24). Moreover, the substantial increase in RNA mass of metazoans, particularly mammalian ribosomes, compared to yeast and protists, has resulted in the presence of two additional RNA layers (Fig. 5e, f): a rigid inner layer, resulting from multiple RNA–RNA tertiary interactions, followed by a flexible outer layer, arising from helical insertions and extensions of the rRNA expansion segments. The observed participation

of rRNA expansion segments in new intersubunit bridges or in the coordination of ribosomal ligands calls for further analysis of their functional significance in the complex environment of the eukaryotic cell.

METHODS SUMMARY

Drosophila and human 80S ribosomes were purified by sucrose density centrifugation from embryo extracts and peripheral blood mononuclear cells, respectively. For cryo-EM, ribosomes were vitrified and data were collected on a Titan Krios EM (FEI Company). Single-particle analysis and three-dimensional reconstruction were carried out using the SPIDER software package[43]. Ribosomal RNA was modelled using S2S[44] and Assemble[45]. Protein models were generated using Modeller[46].

Full Methods and any associated references are available in the online version of the paper.

Received 9 January; accepted 19 March 2013.

1. Schmeing, T. M. & Ramakrishnan, V. What recent ribosome structures have revealed about the mechanism of translation. *Nature* **461,** 1234–1242 (2009).
2. Wilson, D. N. & Cate, J. H. D. The structure and function of the eukaryotic ribosome. *Cold Spring Harb. Perspect. Biol.* **4,** a011536 (2012).
3. Klinge, S., Voigts-Hoffmann, F., Leibundgut, M. & Ban, N. Atomic structures of the eukaryotic ribosome. *Trends Biochem. Sci.* **37,** 189–198 (2012).
4. Melnikov, S. *et al.* One core, two shells: bacterial and eukaryotic ribosomes. *Nature Struct. Mol. Biol.* **19,** 560–567 (2012).
5. Taylor, D. J. *et al.* Comprehensive molecular structure of the eukaryotic ribosome. *Structure* **17,** 1591–1604 (2009).
6. Armache, J. P. *et al.* Cryo-EM structure and rRNA model of a translating eukaryotic 80S ribosome at 5.5-A resolution. *Proc. Natl Acad. Sci. USA* **107,** 19748–19753 (2010).
7. Armache, J. P. *et al.* Localization of eukaryote-specific ribosomal proteins in a 5.5-A cryo-EM map of the 80S eukaryotic ribosome. *Proc. Natl Acad. Sci. USA* **107,** 19754–19759 (2010).
8. Rabl, J., Leibundgut, M., Ataide, S. F., Haag, A. & Ban, N. Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1. *Science* **331,** 730–736 (2011).
9. Klinge, S., Voigts-Hoffmann, F., Leibundgut, M., Arpagaus, S. & Ban, N. Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6. *Science* **334,** 941–948 (2011).
10. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 A resolution. *Science* **334,** 1524–1529 (2011).
11. Dube, P. *et al.* Correlation of the expansion segments in mammalian rRNA with the fine structure of the 80 S ribosome; a cryoelectron microscopic reconstruction of the rabbit reticulocyte ribosome at 21 A resolution. *J. Mol. Biol.* **279,** 403–421 (1998).
12. Spahn, C. M. *et al.* Cryo-EM visualization of a viral internal ribosome entry site bound to human ribosomes; the IRES functions as an RNA-based translation factor. *Cell* **118,** 465–475 (2004).
13. Boehringer, D., Thermann, R., Ostareck-Lederer, A., Lewis, J. D. & Stark, H. Structure of the hepatitis C virus IRES bound to the human 80S ribosome: remodeling of the HCV IRES. *Structure* **13,** 1695–1706 (2005).
14. Chandramouli, P. *et al.* Structure of the mammalian 80S ribosome at 8.7 A resolution. *Structure* **16,** 535–548 (2008).
15. Ruvinsky, I. & Meyuhas, O. Ribosomal protein S6 phosphorylation: from protein synthesis to cell size. *Trends Biochem. Sci.* **31,** 342–348 (2006).
16. Koyama, Y., Katagiri, S., Hanai, S., Uchida, K. & Miwa, M. Poly(ADP-ribose) polymerase interacts with novel *Drosophila* ribosomal proteins, L22 and l23a, with unique histone-like amino-terminal extensions. *Gene* **226,** 339–345 (1999).
17. Ramakrishnan, V. Histone structure and the organization of the nucleosome. *Annu. Rev. Biophys. Biomol. Struct.* **26,** 83–112 (1997).
18. Taylor, D. J. *et al.* Structures of modified eEF2 80S ribosome complexes reveal the role of GTP hydrolysis in translocation. *EMBO J.* **26,** 2421–2431 (2007).
19. Harms, J. M. *et al.* Translational regulation via L11: molecular switches on the ribosome turned on and off by thiostrepton and micrococcin. *Mol. Cell* **30,** 26–38 (2008).
20. Gao, Y. G. *et al.* The structure of the ribosome with elongation factor G trapped in the posttranslocational state. *Science* **326,** 694–699 (2009).
21. Dever, T. E. & Green, R. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **4,** a013706 (2012).
22. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74,** 129–177 (2005).
23. Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M. & Yusupova, G. A new understanding of the decoding principle on the ribosome. *Nature* **484,** 256–259 (2012).
24. Lu, H., Li, W., Noble, W. S., Payan, D. & Anderson, D. C. Riboproteomics of the hepatitis C virus internal ribosomal entry site. *J. Proteome Res.* **3,** 949–957 (2004).
25. Spahn, C. M. *et al.* Hepatitis C virus IRES RNA-induced changes in the conformation of the 40s ribosomal subunit. *Science* **291,** 1959–1962 (2001).
26. Balagopal, V. & Parker, R. Stm1 modulates translation after 80S formation in *Saccharomyces cerevisiae. RNA* **17,** 835–842 (2011).
27. Gerbi, S. A. in *Ribosomal RNA—Structure, Evolution, Processing, and Function in Protein Synthesis* (eds Zimmermann, R. A. & Dahlberg, A. E.) 71–87 (CRC Press, 1996).
28. Haga, J. Y., Hamilton, M. G. & Petermann, M. L. Electron microscopic observations on the large subunit of the rat liver ribosome. *J. Cell Biol.* **47,** 211–221 (1970).
29. Cannone, J. J. *et al.* The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3,** 2 (2002).
30. Fields, D. S. & Gutell, R. R. An analysis of large rRNA sequences folded by a thermodynamic method. *Fold. Des.* **1,** 419–430 (1996).
31. Alkemar, G. & Nygard, O. Probing the secondary structure of expansion segment ES6 in 18S ribosomal RNA. *Biochemistry* **45,** 8067–8078 (2006).
32. Andersen, C. B. *et al.* Structure of eEF3 and the mechanism of transfer RNA release from the E-site. *Nature* **443,** 663–668 (2006).
33. Srivastava, S., Verschoor, A. & Frank, J. Eukaryotic initiation factor-3 does not prevent association through physical blockage of the ribosomal subunit-subunit interface. *J. Mol. Biol.* **226,** 301–304 (1992).
34. Siridechadilok, B., Fraser, C. S., Hall, R. J., Doudna, J. A. & Nogales, E. Structural roles for human translation factor eIF3 in initiation of protein synthesis. *Science* **310,** 1513–1515 (2005).
35. Yu, Y., Abaeva, I. S., Marintchev, A., Pestova, T. V. & Hellen, C. U. Common conformational changes induced in type 2 picornavirus IRESs by cognate trans-acting factors. *Nucleic Acids Res.* **39,** 4851–4865 (2011).
36. Beckmann, R. *et al.* Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107,** 361–372 (2001).
37. Becker, T. *et al.* Structure of monomeric yeast and mammalian Sec61 complexes interacting with the translating ribosome. *Science* **326,** 1369–1373 (2009).
38. Sweeney, R., Chen, L. H. & Yao, M. C. An rRNA variable region has an evolutionarily conserved essential role despite sequence divergence. *Mol. Cell. Biol.* **14,** 4203–4215 (1994).
39. Bradatsch, B. *et al.* Structure of the pre-60S ribosomal subunit with nuclear export factor Arx1 bound at the exit tunnel. *Nature Struct. Mol. Biol.* **19,** 1234–1241 (2012).
40. Greber, B. J., Boehringer, D., Montellese, C. & Ban, N. Cryo-EM structures of Arx1 and maturation factors Rei1 and Jjj1 bound to the 60S ribosomal subunit. *Nature Struct. Mol. Biol.* **19,** 1228–1233 (2012).
41. Leidig, C. *et al.* Structural characterization of a eukaryotic chaperone—the ribosome-associated complex. *Nature Struct. Mol. Biol.* **20,** 23–28 (2013).
42. Blau, M. *et al.* ERj1p uses a universal ribosomal adaptor site to coordinate the 80S ribosome at the membrane. *Nature Struct. Mol. Biol.* **12,** 1015–1016 (2005).
43. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116,** 190–199 (1996).
44. Jossinet, F. & Westhof, E. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* **21,** 3320–3321 (2005).
45. Jossinet, F., Ludwig, T. E. & Westhof, E. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics* **26,** 2057–2059 (2010).
46. Eswar, N., Eramian, D., Webb, B., Shen, M. Y. & Sali, A. Protein structure modeling with MODELLER. *Methods Mol. Biol.* **426,** 145–159 (2008).
47. Jenner, L., Demeshkina, N., Yusupova, G. & Yusupov, M. Structural rearrangements of the ribosome at the tRNA proofreading step. *Nature Struct. Mol. Biol.* **17,** 1072–1078 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Ungewickell for assistance with cryo-EM data collection and P. Palluch for preparation of peripheral blood mononucleic cells. We thank M. Yusupov, A. Ben-Shem, N. Garreau de Loubresse and S. Melnikov for sharing *S. cerevisiae* X-ray data before publication. We thank P. Becker for access to his fly facility and help with embryo collection, and V. Márquez, T. Fröhlich, G. Arnold, I. Forné and A. Imhof for mass-spectrometry analysis. This research was supported by grants from the Deutsche Forschungsgemeinschaft SFB594, SFB646 and GRK 1721 (to R.B.), and FOR1805 (to R.B. and D.N.W.). D.N.W. is supported by the European Molecular Biology Organization (EMBO) young investigator program. This work was supported by a European Research Council (ERC) Advanced Grant (to R.B.).

Author Contributions A.M.A. prepared *D. melanogaster* embryo extracts, purified *D. melanogaster* and *H. sapiens* ribosome samples, carried out mass-spectrometry analysis of *H. sapiens* ribosomes and prepared the figures; A.M.A. and J.-P.A. contributed blood, processed cryo-EM data and built atomic models; O.B. carried out cryo-EM data collection; M.H. performed deconvolution and sharpening on electron density maps; M.S. designed experiments for blood collection and peripheral-blood-mononuclear-cell preparations for human ribosome purification; A.M.A., J.-P.A., D.N.W. and R.B. interpreted results and wrote the manuscript. D.N.W. and R.B. designed research and supervised the project.

Author Information Coordinates of the atomic models have been deposited in the Protein Data Bank with accession numbers 3J38, 3J39, 3J3C and 3J3E for *Drosophila* 80S ribosomes and 3J3A, 3J3B, 3J3D and 3J3F for human 80S ribosomes. Full models can be obtained from the database of aligned ribosomal complexes (DARC) site (http://darcsite.genzentrum.lmu.de/darc/). Electron-microscopy maps of the *Drosophila* and human ribosomes have been deposited in the EM Data Bank under the accession codes EMD-5591 and EMD-5592, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.B. (beckmann@lmb.uni-muenchen.de).

METHODS

Purification of 80S ribosomes from *D. melanogaster*. Extracts from *D. melanogaster* embryos were prepared as described previously[48], and incubated under high-salt conditions (20 mM HEPES, pH 7.4, 500 mM potassium acetate (KOAc), 25 mM magnesium acetate ($Mg(OAc)_2$), 1 mM dithiotreitol (DTT), 0.5 mM phenylmethylsulfonylflourid (PMSF), 0.2 units per μl anti-RNase (Ambion)) with 0.5 mM puromycin for 15 min on ice, then for 10 min at 20 °C. Ribosomes were pelleted through a high-salt sucrose cushion (1 M sucrose, 500 mM KOAc, 25 mM $Mg(OAc)_2$, 1 mM DTT, 0.5 mM PMSF) at 355,040*g* (TLA120.2, Beckman-Coulter) for 60 min. The ribosomal pellet was suspended in buffer A (20 mM HEPES, pH 7.4, 100 mM KOAc, 5 mM $Mg(OAc)_2$, 1 mM DTT, 0.5 mM PMSF) with 125 mM sucrose. Ribosomes were purified further by centrifugation through a linear sucrose density gradient (10–40% sucrose in buffer A) at 202,048g (SW-40 Ti, Beckman Coulter) for 3 h at 4 °C. Fractions were collected using a Gradient Station (Biocomp) with an Econo UV Monitor (Biorad) and a FC203B Fraction Collector (Gilson). Ribosomes were pelleted from suitable fractions by centrifugation at 385,840*g* (TLA110, Beckman-Coulter) for 75 min. The pellet was suspended in buffer B (20 mM HEPES, pH 7.4, 50 mM KOAc, 2.5 mM $Mg(OAc)_2$, 100 mM sucrose, 1 mM DTT, 0.5 mM PMSF).

Purification of human 80S ribosomes. Mononuclear cells were prepared from human peripheral blood by ficoll-hypaque density-gradient centrifugation[49]. Cell pellets were suspended in lysis buffer (20 mM HEPES, pH 7.4, 100 mM KOAc, 7.5 mM $Mg(OAc)_2$, 1 mM DTT) with 1× Complete EDTA-free Protease Inhibitor cocktail (Roche) and lysed by repeated freeze and thaw cycles. Cell debris was removed by centrifugation for 20 min at 20,000*g* at 4 °C. Ribosomes were purified from the lysate after high salt and puromycin treatment as described above.

Electron microscopy and image processing. Samples were applied to 2-nm precoated R3/3 holey carbon supported copper grids (Quantifoil), vitrified using a Vitrobot Mark IV (FEI Company) and visualized on a Titan Krios transmission electron microscope (TEM) (FEI Company) under low-dose conditions (20e^- per $Å^2$) at a nominal magnification of ×75,000 with a nominal defocus between −1.0 and −3.5 μm. Data were collected using the semi-automated software EM-TOOLS (TVIPS GmbH) as described[50]. Contrast-transfer functions were determined using CTFFIND[51] and recorded images were manually inspected for good areas and power-spectra quality. Data were processed further using the SPIDER software package[43], in combination with an automated workflow as described previously[50].

The *D. melanogaster* 80S ribosome data set was collected at 300 keV at a magnification of ×128,200 at the plane of the charge-coupled device (CCD) using an Eagle 4k CCD camera (FEI Company, 4,096 × 4,096 pixels, 15-μm pixels, 5 s per full frame) resulting in an image pixel size of 1.17 Å on the object scale. The total data set consisted of 317,000 particles that entered a second round of selection using a machine-learning algorithm (MAPPOS[72]) that detects non-particles as described previously[50]. This procedure resulted in a cleaned data set of 288,000 (90.9%) particles that were used for the initial alignment. An empty yeast 80S ribosome structure was used as a reference. The data set was sorted[37,52] according to the presence of eEF2. The final (eEF2 and E-tRNA bound) data set contained 134,500 particles (42.4%) and reached a resolution of 6.0 Å after several rounds of refinement.

The *H. sapiens* 80S ribosome data set was collected at 200 keV at a magnification of ×148,721 at the plane of the CCD using a TemCam-F416 CMOS CCD camera (TVIPS GmbH, 4,096 × 4,096 pixels, 15.6-μm pixels, 1 s per full frame), resulting in a pixel size of 1.0489 Å on the object scale. Four separate data collections were used, of which the first (650,000 particles) was carried out using a normal field emission gun (FEG), whereas the remaining three (2.1-million particles) were collected with an X-FEG module (FEI Company) as the electron source. The collected data were initially aligned to a *Triticum aestivum* ribosome[6]. After a few rounds of refinement the data set was sorted[37,52], resulting in two maps representing stable conformations: a non-rotated 80S ribosome with E-tRNA, and a rotated 80S ribosome containing eEF2, SERPB1 and E-tRNA. The complete data were re-aligned using the best-resolved output from the previous refinement attempt (rotated 80S + eEF2 + SERPB1 + E-tRNA). After many rounds of refinement, re-sorting and application of a non-negative deconvolution and sharpening process[53], we arrived at a final average resolution of 5.4 Å from 343,343 particles. Local resolution was computed within a softened sphere (radius of 22 Å) at each voxel, as described previously[54], using the fourier shell correlation (FSC) of two reconstructions; from 50% of the particles and then the other 50%.

Ribosomal RNA modelling. *H. sapiens* 18S, 5S, 28S and 5.8S rRNA sequences were taken from GeneBank entries X03205 and V00589 and RefSeq accession numbers NR_003287 and NR_046235, respectively[55,56]. *D. melanogaster* sequences for the 18S, 28S, 2S, 5.8S and 5S rRNAs were obtained from GeneBank accessions M21017 and M25016, respectively[57,58], in combination with a revised 28S sequence for nucleotides 221–245 (H19 and H20), which are missing in the original sequence[59]. Structure-based sequence alignments of the conserved rRNA core were constructed using Sequence to Structure (S2S)[44] based on the X-ray structure of the 80S ribosome from *S. cerevisiae* (Protein Data Bank (PDB) accession codes 3O58 and 3O2Z)[10]. For the L1-stalk region (H76–H78) the corresponding structure of *Escherichia coli* (PDB accession 3R8S)[60] was used as template in a separate S2S alignment. All remaining parts of the rRNA were built *de novo* using Assemble[45], guided by features of the electron-density and secondary-structure predictions from RNAfold[61], in the main as described previously[6]. Secondary structures of large rRNA parts were predicted in smaller pieces and then by inspection of the corresponding electron-density map and subsequent model building. This generated new rRNA boundaries that were used as starting points for secondary-structure predictions of the following sequences. The iterative process resulted in the identification of simple, un-branched helical folds for the flexible human rRNA arms and enabled us to build complete molecular models of the human and *Drosophila* rRNA. The models were adjusted according to features of the electron density using Assemble[45], molecular dynamic flexible fitting (MDFF)[62] in visual molecular dynamics (VMD)[63] and Coot[64]. The reliability of the molecular model for the rRNA is indicated using the *b*-factor values within the PDB file; more reliably modelled regions have a lower *b*-factor.

Ribosomal protein modelling. Owing to the availability of ribosomal 40S[8] and 60S structures[9] from *T. thermophila* and ribosomal 80S structures from *S. cerevisiae*[10], proteins were screened for the best fit into the *D. melanogaster* and *H. sapiens* densities. Protein multi-alignment was carried out using Jalview[65], ClustalW[66], ClustalOmega[67] and Mafft[68]. Results were extracted and Modeller[46] was used to create the initial models. Using UCSF Chimera[69] and Coot[64], they were fitted rigidly and adjusted into the densities. Subsequently, the remaining additional densities were analysed, assigned to specific protein entities and, in conjunction with secondary structure predictions, the models were extended. Furthermore, VMD[63], MDFF[62] and Coot were used to fix the clashes. The reliability of the molecular model for the ribosomal proteins is indicated using the *b*-factor values within the PDB file. More reliably modelled regions have a lower *b*-factor.

Mass-spectrometry analysis. For the *Drosophila* ribosome, proteins were extracted by acetic acid according to a previous study[70], and subsequent liquid chromatography tandem mass spectrometry (LC–MS/MS) analysis of the protein samples was carried out as described previously[71]. For the human ribosome sample, proteins were reduced, alkylated and digested with trypsin in solution before desalting and subsequent LC–MS/MS analysis using a LTQ Orbitrap XL (Thermo Scientific) mass spectrometer. MS/MS data were searched with Mascot (Matrix Science) using the SwissProt 2011.02 database and the following parameters: enzyme, trypsin; fixed modification, carbamidomethyl; variable modification, oxidation; peptide-mass tolerance, 10 p.p.m.; fragment mass tolerance, 0.8 Da; and up to two missed cleavages allowed.

Figure preparation. Figures showing electron densities and atomic models were generated using UCSF Chimera[69].

48. Gebauer, F., Corona, D. F., Preiss, T., Becker, P. B. & Hentze, M. W. Translational control of dosage compensation in *Drosophila* by Sex-lethal: cooperative silencing via the 5′ and 3′ UTRs of msl-2 mRNA is independent of the poly(A) tail. *EMBO J.* **18,** 6146–6154 (1999).
49. Fuss, I. J., Kanof, M. E., Smith, P. D. & Zola, H. Isolation of whole mononuclear cells from peripheral blood and cord blood. *Curr. Protoc. Immunol.* **85,** 7.1.1–7.1.8 (2009).
50. Becker, T. *et al.* Structural basis of highly conserved ribosome recycling in eukaryotes and archaea. *Nature* **482,** 501–506 (2012).
51. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
52. Becker, T. *et al.* Structure of the no-go mRNA decay complex Dom34-Hbs1 bound to a stalled 80S ribosome. *Nature Struct. Mol. Biol.* **18,** 715–720 (2011).
53. Hirsch, M., Scholkopf, B. & Habeck, M. A blind deconvolution approach for improving the resolution of cryo-EM density maps. *J. Comput. Biol.* **18,** 335–346 (2011).
54. Lasker, K. *et al.* Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc. Natl Acad. Sci. USA* **109,** 1380–1387 (2012).
55. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
56. Maden, B. E. *et al.* Clones of human ribosomal DNA containing the complete 18 S-rRNA and 28 S-rRNA genes. Characterization, a detailed map of the human ribosomal transcription unit and diversity among clones. *Biochem. J.* **246,** 519–527 (1987).
57. Tautz, D., Hancock, J. M., Webb, D. A., Tautz, C. & Dover, G. A. Complete sequences of the rRNA genes of *Drosophila melanogaster*. *Mol. Biol. Evol.* **5,** 366–376 (1988).
58. Thompson, J. F., Wegnez, M. R. & Hearst, J. E. Determination of the secondary structure of *Drosophila melanogaster* 5 S RNA by hydroxymethyltrimethylpsoralen crosslinking. *J. Mol. Biol.* **147,** 417–436 (1981).
59. Rousset, F., Pelandakis, M. & Solignac, M. Evolution of compensatory substitutions through G.U intermediate state in *Drosophila* rRNA. *Proc. Natl Acad. Sci. USA* **88,** 10032–10036 (1991).
60. Dunkle, J. A. *et al.* Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science* **332,** 981–984 (2011).

61. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31,** 3429–3431 (2003).
62. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16,** 673–683 (2008).
63. Humphrey, W., Dalke, A. & Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).
64. Emsley, P. & Cowtan, K. Coot: model-Building Tools for Molecular Graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132 (2004).
65. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).
66. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
67. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
68. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33,** 511–518 (2005).
69. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
70. Nierhaus, K. H. & Dohme, F. Total reconstitution of functionally active 50S ribosomal subunits from *E. coli. Proc. Natl Acad. Sci. USA* **71,** 4713–4717 (1974).
71. Márquez, V. *et al.* Proteomic characterization of archaeal ribosomes reveals the presence of novel archaeal-specific ribosomal proteins. *J. Mol. Biol.* **405,** 1215–1232 (2011).
72. Norousi, R. *et al.* Automated post-picking using MAPPOS improves particle image detection from cryo-EM micrographs. *J. Struct. Biol.* http://dx.doi.org/10.1016/j.jsb.2013.02.008 (2013).

LETTER

doi:10.1038/nature12016

Heralded entanglement between solid-state qubits separated by three metres

H. Bernien[1], B. Hensen[1], W. Pfaff[1], G. Koolstra[1], M. S. Blok[1], L. Robledo[1], T. H. Taminiau[1], M. Markham[2], D. J. Twitchen[2], L. Childress[3] & R. Hanson[1]

Quantum entanglement between spatially separated objects is one of the most intriguing phenomena in physics. The outcomes of independent measurements on entangled objects show correlations that cannot be explained by classical physics. As well as being of fundamental interest, entanglement is a unique resource for quantum information processing and communication. Entangled quantum bits (qubits) can be used to share private information or implement quantum logical gates[1,2]. Such capabilities are particularly useful when the entangled qubits are spatially separated[3–5], providing the opportunity to create highly connected quantum networks[6] or extend quantum cryptography to long distances[7,8]. Here we report entanglement of two electron spin qubits in diamond with a spatial separation of three metres. We establish this entanglement using a robust protocol based on creation of spin–photon entanglement at each location and a subsequent joint measurement of the photons. Detection of the photons heralds the projection of the spin qubits onto an entangled state. We verify the resulting non-local quantum correlations by performing single-shot readout[9] on the qubits in different bases. The long-distance entanglement reported here can be combined with recently achieved initialization, readout and entanglement operations[9–13] on local long-lived nuclear spin registers, paving the way for deterministic long-distance teleportation, quantum repeaters and extended quantum networks.

A quantum network can be constructed by using entanglement to connect local processing nodes, each containing a register of well-controlled and long-lived qubits[6]. Solids are an attractive platform for such registers, as the use of nanofabrication and material design may enable well-controlled and scalable qubit systems[14]. The potential impact of quantum networks on science and technology has recently spurred research efforts towards generating entangled states of distant solid-state qubits[15–21].

A prime candidate for a solid-state quantum register is the nitrogen–vacancy (NV) defect centre in diamond. The NV centre combines a long-lived electronic spin ($S = 1$) with a robust optical interface, enabling measurement and high-fidelity control of the spin qubit[15,22–24]. Furthermore, the NV electron spin can be used to access and manipulate nearby nuclear spins[9–13,25], thereby forming a multi-qubit register. To use such registers in a quantum network requires a mechanism to coherently connect remote NV centres.

Here we demonstrate the generation of entanglement between NV centre spin qubits in distant set-ups. We achieve this by combining recently established spin initialization and single-shot readout techniques[9] with efficient resonant optical detection and feedback-based control over the optical transitions, all in a single experiment and executed with high fidelity. These results put solid-state qubits on a par with trapped atomic qubits[3–5] as highly promising candidates for implementing quantum networks.

Our experiment makes use of two NV spin qubits located in independent low-temperature set-ups separated by 3 m (Fig. 1a). We encode the qubit basis states $|\uparrow\rangle$ and $|\downarrow\rangle$ in the NV spin sublevels $m_S = 0$ and $m_S = -1$, respectively. Each qubit can be independently read out by detecting spin-dependent fluorescence in the NV phonon sideband (non-resonant detection)[9]. The qubits are individually controlled with microwave pulses applied to on-chip striplines[23]. Quantum states encoded in the qubits are extremely long-lived: using dynamical decoupling techniques[23], we obtain a coherence time exceeding 10 ms (Fig. 1b), which is the longest coherence time measured so far for a single electron spin in a solid.

We generate and herald entanglement between these distant qubits by detecting the resonance fluorescence of the NV centres. The specific entanglement protocol we use is based on the proposal of ref. 26, and is schematically drawn in Fig. 1c. Both centres NV A and NV B are initially prepared in a superposition $1/\sqrt{2}(|\uparrow\rangle + |\downarrow\rangle)$. Next, each NV centre is excited by a short laser pulse that is resonant with the $|\uparrow\rangle$ to $|e\rangle$ transition, where $|e\rangle$ is an optically excited state with the same spin projection as $|\uparrow\rangle$. Spontaneous emission locally entangles the qubit and photon number, leaving each set-up in the state $1/\sqrt{2}(|\uparrow 1\rangle + |\downarrow 0\rangle)$, where 1 (0) denotes the presence (absence) of an emitted photon; the joint qubit–photon state of both set-ups is then described by $1/2(|\uparrow_A\uparrow_B\rangle|1_A1_B\rangle + |\downarrow_A\downarrow_B\rangle|0_A0_B\rangle + |\uparrow_A\downarrow_B\rangle|1_A0_B\rangle + |\downarrow_A\uparrow_B\rangle|0_A1_B\rangle)$. The two photon modes, A and B, are directed to the input ports of a beamsplitter (see Fig. 1a), so that fluorescence observed in an output port could have originated from either NV centre. If the photons emitted by the two NV centres are indistinguishable, detection of precisely one photon on an output port would correspond to measuring the photon state $1/\sqrt{2}(|1_A0_B\rangle \pm e^{-i\varphi}|0_A1_B\rangle)$ (where φ is a phase that depends on the optical path length). Such a detection event would thereby project the qubits onto the maximally entangled state $|\psi\rangle = 1/\sqrt{2}(|\uparrow_A\downarrow_B\rangle \pm e^{-i\varphi}|\downarrow_A\uparrow_B\rangle)$.

Any realistic experiment, however, suffers from photon loss and imperfect detector efficiency; detection of a single photon is thus also consistent with creation of the state $|\uparrow\uparrow\rangle$. To eliminate this possibility, both qubits are flipped and optically excited for a second time. Because $|\uparrow\uparrow\rangle$ is flipped to $|\downarrow\downarrow\rangle$, no photons are emitted in the second round for this state. In contrast, the states $|\psi\rangle$ will again yield a single photon. Detection of a photon in both rounds thus heralds the generation of an entangled state. The second round not only renders the protocol robust against photon loss, but it also changes φ into a global phase, making the protocol insensitive to the optical path length difference[26] (see Supplementary Information). Furthermore, flipping the qubits provides a refocusing mechanism that counteracts spin dephasing during entanglement generation. The final state is one of two Bell states $|\Psi^{\pm}\rangle = 1/\sqrt{2}(|\uparrow_A\downarrow_B\rangle \pm |\downarrow_A\uparrow_B\rangle)$, with the sign depending on whether the same detector (+) or different detectors (−) clicked in the two rounds.

A key challenge for generating remote entanglement with solid-state qubits is obtaining a large flux of indistinguishable photons, in part because local strain in the host lattice can induce large variations in photon frequency. The optical excitation spectra of the NV centres

[1]Kavli Institute of Nanoscience Delft, Delft University of Technology, PO Box 5046, 2600 GA Delft, The Netherlands. [2]Element Six Ltd, Kings Ride Park, Ascot, Berkshire SL5 8BP, UK. [3]McGill University Department of Physics, 3600 Rue University, Montreal, Quebec H3A 2T8, Canada.

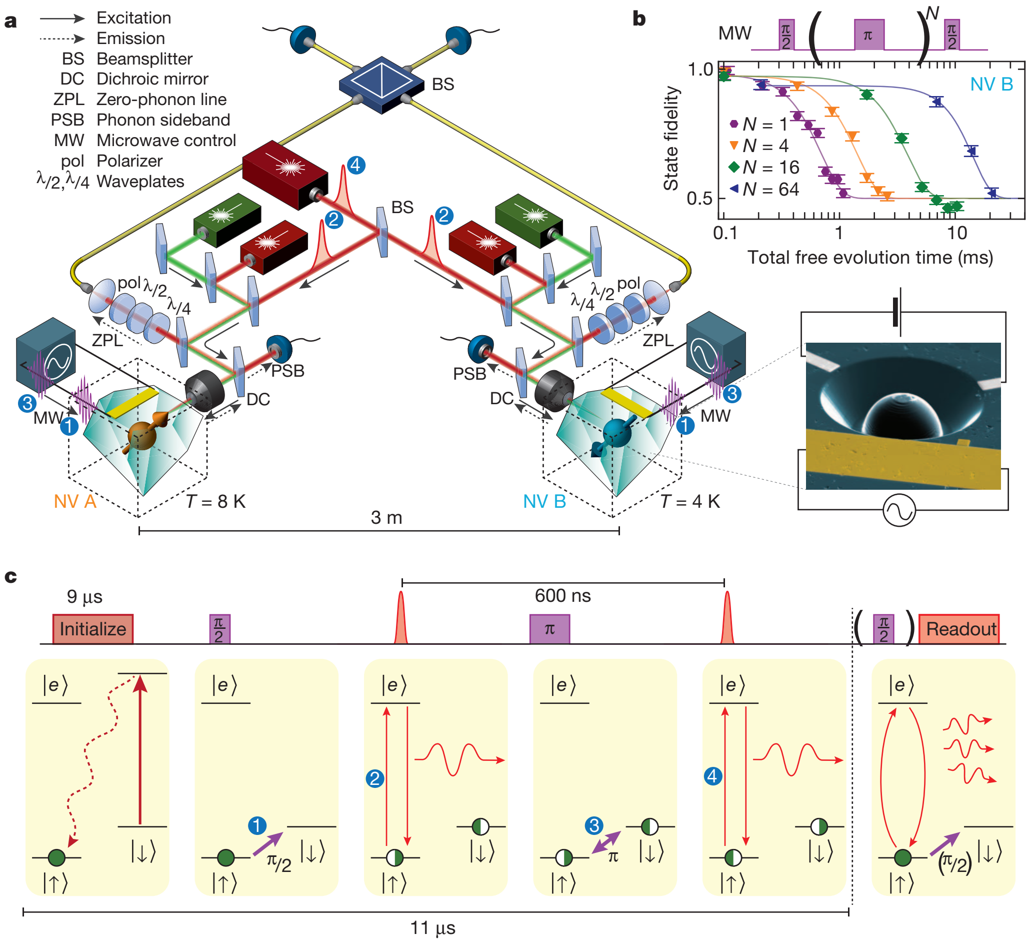


Figure 1 | Experimental set-up and protocol for generating long-distance entanglement between two solid-state spin qubits. a, Experimental set-up. Each nitrogen–vacancy (NV) centre resides in a synthetic ultrapure diamond oriented in the <111> direction. The two diamonds are located in two independent low-temperature confocal microscope set-ups separated by 3 m. The NV centres can be individually excited resonantly by red lasers and off-resonantly by a green laser. The emission (dashed arrows) is spectrally separated into an off-resonant part (phonon sideband, PSB) and a resonant part (zero-phonon line, ZPL). The PSB emission is used for independent single-shot readout of the spin qubits[9]. The ZPL photons from the two NV centres are overlapped on a fibre-coupled beamsplitter. Microwave pulses for spin control are applied via on-chip microwave striplines. An applied magnetic field of 17.5 G splits the $m_S = \pm 1$ levels in energy. The optical frequencies of NV B are tuned by a d.c. electric field applied to the gate electrodes (inset, scanning electron microscope image of a similar device). To enhance the collection efficiency, solid immersion lenses have been milled around the two NV centres[9]. **b**, The coherence of the NV B spin qubit as a function of total free evolution time t_{FE} during an N-pulse dynamical decoupling sequence[23]. Curves are fitted to $A\exp[-(t_{FE}/T_{coh})^3] + 0.5$. For $N = 64$ we find $T_{coh} = 14.3 \pm 0.4$ ms. Error bars are 2 s.e. **c**, Entanglement protocol (details in main text), illustrating the pulse sequence applied simultaneously to both NV centres. Both NV centres are initially prepared in a superposition $1/\sqrt{2}(|\uparrow\rangle + |\downarrow\rangle)$. A short 2 ns spin-selective resonant laser pulse creates spin–photon entanglement $1/\sqrt{2}(|\uparrow 1\rangle + |\downarrow 0\rangle)$. The photons are overlapped on the beamsplitter and detected in the two output ports. Both spins are then flipped, and the NV centres are excited a second time. The detection of one photon in each excitation round heralds the entanglement and triggers individual spin readout.

(Fig. 2a) display sharp spin-selective transitions. Here we use the E_y transition (spin projection $m_S = 0$) in the entangling protocol and for qubit readout; we use the A_1 transition for fast optical pumping into $|\uparrow\rangle$ (ref. 9). Owing to different strain in the two diamonds, the frequencies of the E_y transitions differ by 3.5 GHz, more than 100 linewidths. By applying a voltage to an on-chip electrode (Fig. 1a inset), we tune the optical transition frequencies of one centre (NV B) through the d.c. Stark effect[18,27] and bring the E_y transitions of the two NV centres into resonance (Fig. 2a bottom).

Charge fluctuations near the NV centre also affect the optical frequencies. To counteract photo-ionization, we need to regularly apply a green laser pulse to repump the NV centre into the desired charge state. This repump pulse changes the local electrostatic environment, leading to jumps of several linewidths in the optical transition frequencies[28]. To overcome these effects, we only initiate an experiment if the number of photons collected during a two-laser probe stage (Fig. 2b) exceeds a threshold, thereby ensuring that the NV centre's optical transitions are on resonance with the lasers. The preparation procedure markedly improves the observed optical coherence: as the probe threshold is increased, optical Rabi oscillations persist for longer times (see Fig. 2b). For high thresholds, the optical damping time saturates around the value expected for a lifetime-limited linewidth[28], indicating that the effect of spectral jumps induced by the repump laser is strongly mitigated.

Besides photon indistinguishability, successful execution of the protocol also requires that the detection probability of resonantly emitted photons exceed that of scattered laser photons and of detector dark counts. This is particularly demanding for NV centres, because only about 3% of their emission is in the zero-phonon line and useful for the protocol. To minimize detection of laser photons, we use both a cross-polarized excitation–detection scheme (Fig. 2c inset) and a detection time filter that exploits the difference between the length of the laser pulse (2 ns) and the NV centre's excited-state lifetime (12 ns; Fig. 2c). For a typical detection window used, this reduces the contribution of scattered laser photons to about 1%. Combined with microfabricated solid-immersion lenses for enhanced collection efficiency (Fig. 1a

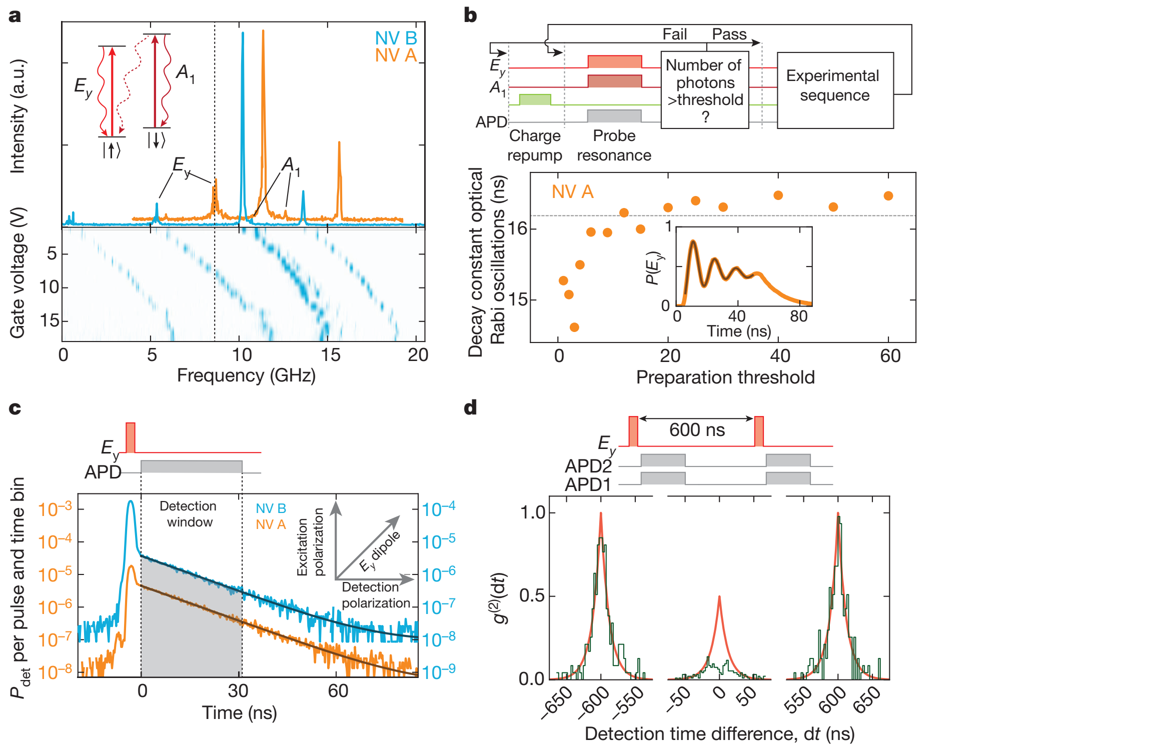


Figure 2 | Generating and detecting indistinguishable photons. **a**, Photoluminescence excitation spectra of NV A and NV B; frequency is given relative to 470.4515 THz. Transitions are labelled according to the symmetry of their excited state. The A_1 transition is used to initialize the NV centre into the $|\uparrow\rangle$ state ($m_S = 0$) and the E_y transition is used for entanglement creation and single-shot readout. By applying a voltage to the gate electrodes of NV B, the E_y transitions are tuned into resonance (dashed line). **b**, Dynamical preparation of charge and optical resonance. Top, preparation protocol. A 10 µs green laser pulse (green line) pumps the NV centre into the desired negative charge state[9]. Next, the optical transition frequencies are probed by simultaneously exciting the E_y and A_1 transitions for 60 µs while counting the number of detected photons. Conditional on passing a certain threshold the experimental sequence is started (preparation successful) or else the protocol is repeated (preparation failed). APD, avalanche photodiode. Bottom, line-narrowing effect of the preparation protocol exemplified by the dependence of the decay time of optical Rabi oscillations on preparation threshold. Dashed line indicates lifetime-limited damping[28]. For the entanglement experiment, we choose a threshold of 45 (20) photons for NV A (NV B). **c**, Resonant optical excitation and detection. The polarization axis of the detection path is aligned perpendicular to the excitation axis. The dipole axis of the E_y transition is oriented in between these two axes (inset). Remaining laser light reflection is time-filtered by defining a photon detection window that starts after the laser pulse. Data are recorded with 256 ps time bins. $P_{\rm det}$, detection probability. **d**, Two-photon quantum interference using resonant excitation and detection. The $g^{(2)}$ correlation function is obtained from all coincidence detection events of APD 1 and APD 2 during the entanglement experiment (see Supplementary Information). The sidepeaks are fitted to an exponential decay; from the fit values, we obtain the expected central peak shape $g^{(2)}_{\perp}$ (red line) for non-interfering photons. The visibility of the interference is given by $(g^{(2)}_{\perp} - g^{(2)})/g^{(2)}_{\perp}$.

inset) and spectral filtering for suppressing non-resonant NV emission, we obtain a detection probability of a resonant NV photon of about 4×10^{-4} per pulse—about 70 times higher than the sum of background contributions.

The degree of photon indistinguishability and background suppression can be obtained directly from the second-order autocorrelation function $g^{(2)}$, which we extract from our entanglement experiment (see Supplementary Information). For fully distinguishable photons, the value of $g^{(2)}$ would reach 0.5 at zero arrival time difference. A strong deviation from this behaviour is observed (Fig. 2d) due to two-photon quantum interference[29] that, for perfectly indistinguishable photons, would make the central peak fully vanish. The remaining coincidences are likely to be caused by (temperature-dependent) phonon-induced transitions between optically excited states[30] in NV A (these transitions are less relevant for NV B because it is at a lower temperature). The visibility of the two-photon interference observed here—(80 ± 5)% for $|dt| < 2.56$ ns—is a significant improvement over previously measured values[18,19] and central to the success of the entangling scheme.

To generate and detect remote entanglement experimentally, we run the following sequence: first, both NV centres are independently prepared into the correct charge state and brought into optical resonance according to the scheme in Fig. 2b. Then we apply the entangling protocol shown in Fig. 1c using a 600 ns delay between the two optical excitation rounds. We repeat the protocol 300 times before we return to the resonance preparation step; this number is a compromise between maximizing the attempt rate and minimizing the probability of NV centre ionization. A fast logic circuit monitors the photon counts in real time and triggers single-shot qubit readout on each set-up whenever entanglement is heralded, that is, whenever a single photon is detected in each round of the protocol. The readout projects each qubit onto the $\{|\uparrow\rangle, |\downarrow\rangle\}$ states (Z-basis), or onto the $\{|\uparrow\rangle \pm |\downarrow\rangle, |\uparrow\rangle \mp |\downarrow\rangle\}$ states (X or $-X$ basis). The latter two are achieved by first rotating the qubit by $\pi/2$ using a microwave pulse before readout. By correlating the resulting single-qubit readout outcomes, we can verify the generation of the desired entangled states. To obtain reliable estimates of the two-qubit state probabilities, we correct the raw data with a maximum-likelihood method for local readout errors. These readout errors are known accurately from regular calibrations performed during the experiment (see Supplementary Information).

Figure 3 shows the obtained correlations. When both qubits are measured along Z (readout basis $\{Z, Z\}$), the states Ψ^+ and Ψ^- (as identified by their different photon signatures) display strongly anti-correlated readout results (odd parity). The coherence of the joint qubit state is revealed by measurements performed in rotated bases ($\{X, X\}$, $\{-X, X\}$), which also exhibit significant correlations. Furthermore, these measurements allow us to distinguish between states Ψ^+ and Ψ^-. For Ψ^+ the $\{X, X\}$ ($\{-X, X\}$) outcomes exhibit even (odd) parity, whereas the Ψ^- state displays the opposite behaviour, as expected. The

observed parities demonstrate that the experiment yields the two desired entangled states.

We calculate a strict lower bound on the state fidelity by combining the measurement results from different bases (see Supplementary Information):

$$F = \langle \Psi^{\pm}|\rho|\Psi^{\pm}\rangle \geq 1/2(P_{\uparrow\downarrow} + P_{\downarrow\uparrow} + C) - \sqrt{(P_{\uparrow\uparrow}P_{\downarrow\downarrow})} \qquad (1)$$

where P_{ij} is the probability for the measurement outcome ij in the $\{Z, Z\}$ basis (that is, the diagonal elements of the density matrix ρ) and C is the contrast between odd and even outcomes in the rotated bases. We find a lower bound of $(69 \pm 5)\%$ for Ψ^- and $(58 \pm 6)\%$ for Ψ^+, and probabilities of 99.98% and 91.8%, respectively, that the state fidelity is above the classical limit of 0.5. These values firmly establish that we have created remote entanglement, and are the main result of this Letter.

The lower bound on the state fidelity given above takes into account the possible presence of coherence within the even-parity subspace $\{|\uparrow\uparrow\rangle, |\downarrow\downarrow\rangle\}$. However, the protocol selects out states with odd parity and therefore this coherence is expected to be absent (see Supplementary Information). To compare the results to the expected value and to account for sources of error, we set the related (square-root) term in equation (1) to zero and obtain for the data in Fig. 3 as best estimate $F = (73 \pm 4)\%$ for Ψ^- and $F = (64 \pm 5)\%$ for Ψ^+.

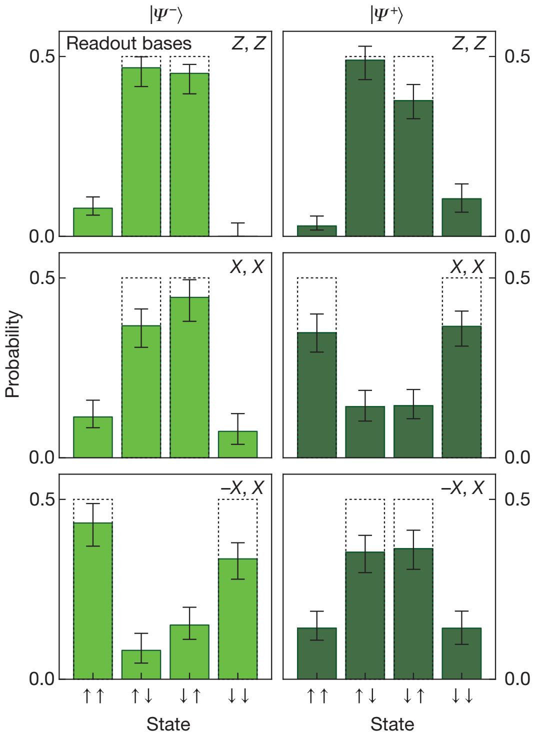


Figure 3 | Verification of entanglement using spin–spin correlations. Each time that entanglement is heralded the spin qubits are individually read out and their results correlated. The readout bases for NV A and NV B can be rotated by individual microwave control (see text). The state probabilities are obtained by a maximum-likelihood estimation on the raw readout results (see Supplementary Information). Error bars depict 68% confidence intervals; dashed lines indicate expected results for perfect state fidelity. Data are obtained from 739 heralding events. For Ψ^-, the detection window in each round is set to 38.4 ns, and the maximum absolute detection time difference $|\delta\tau|$ between the two photons relative to their laser pulses is restricted to 25.6 ns. $\delta\tau = \tau_2 - \tau_1$, where τ_1 is the arrival time of the first photon relative to the first laser pulse and τ_2 the arrival time of the second photon relative to the second laser pulse. For Ψ^+ the second detection window is set to 19.2 ns with $|\delta\tau| < 12.8$ ns, in order to reduce the effect of photo-detector afterpulsing.

Several known error sources contribute to the observed fidelity. Most importantly, imperfect photon indistinguishability reduces the coherence of the state. In Fig. 4a we plot the maximum state fidelity expected from photon interference data (Fig. 2d) together with the measured state fidelities, as a function of the maximum allowed difference in detection time of the two photons relative to their respective laser pulses. We find that the fidelity can be slightly increased by restricting the data to smaller time differences, albeit at the cost of a lower success rate (Fig. 4b).

The fidelity is further decreased by errors in the microwave pulses (estimated at 3.5%), spin initialization (2%), spin decoherence (<1%) and spin flips during the optical excitation (1%) (see Supplementary Information). Moreover, Ψ^+ is affected by afterpulsing, whereby detection of a photon in the first round triggers a fake detector click in the second round. Such afterpulsing leads to a distortion of the correlations (see, for example, the increased probability for $|\downarrow\downarrow\rangle$ in Fig. 3) and thereby a reduction in fidelity for Ψ^+ (see Supplementary Information). Besides these errors that reduce the actual state fidelity, the measured value is also slightly lowered by a conservative estimation for readout errors and by errors in the final microwave $\pi/2$ pulse used for reading out in a rotated basis.

The fidelity of the remote entanglement could be significantly increased in future experiments by further improving photon indistinguishability. This may be achieved by more stringent frequency selection in the resonance initialization step and by working at lower temperatures, which will reduce phonon-mediated excited-state mixing[30]. Also, the microwave errors can be much reduced; for instance, by using isotopically purified diamonds[12] and polarizing the host nitrogen nuclear spin[9].

The success probability of the protocol is given by $P_\Psi = 1/2\ \eta_A\eta_B$. Here η_i is the overall detection efficiency of resonant photons from NV i and the factor 1/2 takes into account cases where the two spins are projected into $|\downarrow\downarrow\rangle$ or $|\uparrow\uparrow\rangle$, which are filtered out by their different photon signature. In the current experiment, we estimate $P_\Psi \approx 10^{-7}$ from the data in Fig. 2c. The entanglement attempt rate is about

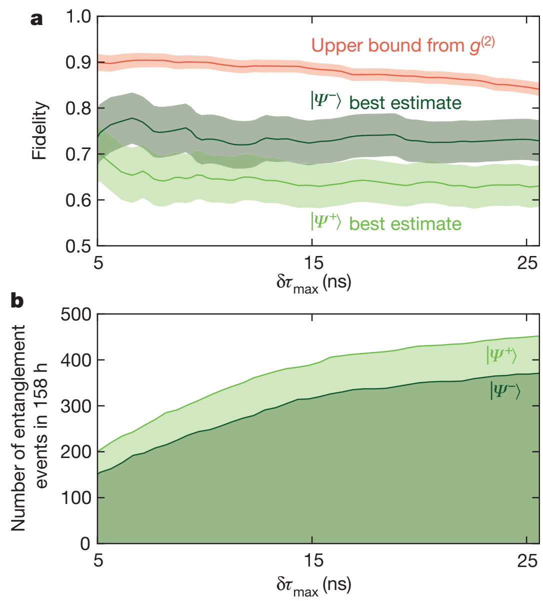


Figure 4 | Dependence of the fidelity and the number of entanglement events on the detection time difference of the photons. a, Upper bound on the state fidelity from photon interference data (see Supplementary Information) and best estimate of the state fidelity from the correlation data as a function of the maximum allowed photon detection time difference ($|\delta\tau| < \delta\tau_{max}$). Detection time windows are chosen as in Fig. 3. Shaded regions indicate 68% confidence intervals. **b**, Number of entanglement events obtained during 158 h as a function of the maximum allowed photon detection time difference, $\delta\tau_{max}$.

20 kHz, yielding one entanglement event per 10 min. This is in good agreement with the 739 entanglement events obtained over a time of 158 h. The use of optical cavities would greatly enhance both the collection efficiency and emission in the zero-phonon line[31] and increase the success rate by several orders of magnitude.

Creation of entanglement between distant spin qubits in diamond, as reported here, opens the door to extending the remarkable properties of NV-based quantum registers towards applications in quantum information science. By transferring entanglement to nuclear spins near each NV centre, a non-local state might be preserved for seconds or longer[12], facilitating the construction of cluster states[2] or quantum repeaters[8]. At the same time, the auxiliary nuclear spin qubits also provide an excellent resource for processing and error correction. When combined with future advances in nanofabricated integrated optics and electronics, the use of electrons and photons as quantum links and nuclear spins for quantum processing and memory offers a compelling route towards realization of solid-state quantum networks.

Received 24 December 2012; accepted 14 February 2013.
Published online 24 April 2013.

1. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, 2000).
2. Raussendorf, R. & Briegel, H. J. A. One-way quantum computer. *Phys. Rev. Lett.* **86,** 5188–5191 (2001).
3. Moehring, D. L. *et al.* Entanglement of single-atom quantum bits at a distance. *Nature* **449,** 68–71 (2007).
4. Ritter, S. *et al.* An elementary quantum network of single atoms in optical cavities. *Nature* **484,** 195–200 (2012).
5. Hofmann, J. *et al.* Heralded entanglement between widely separated atoms. *Science* **337,** 72–75 (2012).
6. Kimble, H. J. The quantum internet. *Nature* **453,** 1023–1030 (2008).
7. Duan, L. M., Lukin, M. D., Cirac, J. I. & Zoller, P. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414,** 413–418 (2001).
8. Childress, L., Taylor, J. M., Sørensen, A. S. & Lukin, M. D. Fault-tolerant quantum communication based on solid-state photon emitters. *Phys. Rev. Lett.* **96,** 070504 (2006).
9. Robledo, L. *et al.* High-fidelity projective read-out of a solid-state spin quantum register. *Nature* **477,** 574–578 (2011).
10. Neumann, P. *et al.* Single-shot readout of a single nuclear spin. *Science* **329,** 542–544 (2010).
11. Neumann, P. *et al.* Multipartite entanglement among single spins in diamond. *Science* **320,** 1326–1329 (2008).
12. Maurer, P. C. *et al.* Room-temperature quantum bit memory exceeding one second. *Science* **336,** 1283–1286 (2012).
13. Pfaff, W. *et al.* Demonstration of entanglement-by-measurement of solid-state qubits. *Nature Phys.* **9,** 29–33 (2013).
14. Ladd, T. D. *et al.* Quantum computers. *Nature* **464,** 45–53 (2010).
15. Togan, E. *et al.* Quantum entanglement between an optical photon and a solid-state spin qubit. *Nature* **466,** 730–734 (2010).
16. Gao, W. B., Fallahi, P., Togan, E., Miguel-Sanchez, J. & Imamoglu, A. Observation of entanglement between a quantum dot spin and a single photon. *Nature* **491,** 426–430 (2012).
17. De Greve, K. *et al.* Quantum-dot spin–photon entanglement via frequency downconversion to telecom wavelength. *Nature* **491,** 421–425 (2012).
18. Bernien, H. *et al.* Two-photon quantum interference from separate nitrogen vacancy centers in diamond. *Phys. Rev. Lett.* **108,** 043604 (2012).
19. Sipahigil, A. *et al.* Quantum interference of single photons from remote nitrogen-vacancy centers in diamond. *Phys. Rev. Lett.* **108,** 143601 (2012).
20. Patel, R. B. *et al.* Two-photon interference of the emission from electrically tunable remote quantum dots. *Nature Photon.* **4,** 632–635 (2010).
21. Flagg, E. B. *et al.* Interference of single photons from two separate semiconductor quantum dots. *Phys. Rev. Lett.* **104,** 137401 (2010).
22. Fuchs, G. D., Dobrovitski, V. V., Toyli, D. M., Heremans, F. J. & Awschalom, D. D. Gigahertz dynamics of a strongly driven single quantum spin. *Science* **326,** 1520–1522 (2009).
23. De Lange, G., Wang, Z. H., Ristè, D., Dobrovitski, V. V. & Hanson, R. Universal dynamical decoupling of a single solid-state spin from a spin bath. *Science* **330,** 60–63 (2010).
24. van der Sar, T. *et al.* Decoherence-protected quantum gates for a hybrid solid-state spin register. *Nature* **484,** 82–86 (2012).
25. Dolde, F. *et al.* Room-temperature entanglement between single defect spins in diamond. *Nature Phys.* **9,** 139–143 (2013).
26. Barrett, S. D. & Kok, P. Efficient high-fidelity quantum computation using matter qubits and linear optics. *Phys. Rev. A* **71,** 060310 (2005).
27. Bassett, L. C., Heremans, F. J., Yale, C. G., Buckley, B. B. & Awschalom, D. D. Electrical tuning of single nitrogen-vacancy center optical transitions enhanced by photoinduced fields. *Phys. Rev. Lett.* **107,** 266403 (2011).
28. Robledo, L., Bernien, H., Van Weperen, I. & Hanson, R. Control and coherence of the optical transition of single nitrogen vacancy centers in diamond. *Phys. Rev. Lett.* **105,** 177403 (2010).
29. Hong, C. K., Ou, Z. Y. & Mandel, L. Measurement of subpicosecond time intervals between two photons by interference. *Phys. Rev. Lett.* **59,** 2044–2046 (1987).
30. Fu, K.-M. C. *et al.* Observation of the dynamic Jahn-Teller effect in the excited states of nitrogen-vacancy centers in diamond. *Phys. Rev. Lett.* **103,** 256404 (2009).
31. Aharonovich, I., Greentree, A. D. & Prawer, S. Diamond photonics. *Nature Photon.* **5,** 397–405 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank F. Jelezko, P. Kok, M. Lukin, J. Morton, E. Togan and L. Vandersypen for discussions and comments, and R. N. Schouten and M. J. Tiggelman for technical assistance. We acknowledge support from the Dutch Organization for Fundamental Research on Matter (FOM), the Netherlands Organization for Scientific Research (NWO), the DARPA QuASAR programme, the EU SOLID, DIAMANT and S3NANO programmes and the European Research Council through a Starting Grant.

Author Contributions H.B., B.H., L.R, L.C. and R.H. designed the experiment. H.B., B.H., W.P., G.K. and M.S.B. performed the experiments. H.B., B.H., W.P., G.K., M.S.B., T.H.T. and R.H. analysed the results. H.B., M.M. and D.J.T. fabricated the devices. H.B., B.H., W.P., M.S.B., L.C. and R.H. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.H. (r.hanson@tudelft.nl).

LETTER

doi:10.1038/nature12081

Optical addressing of an individual erbium ion in silicon

Chunming Yin[1], Milos Rancic[2], Gabriele G. de Boo[1], Nikolas Stavrias[3], Jeffrey C. McCallum[3], Matthew J. Sellars[2] & Sven Rogge[1]

The detection of electron spins associated with single defects in solids is a critical operation for a range of quantum information and measurement applications under development[1–9]. So far, it has been accomplished for only two defect centres in crystalline solids: phosphorus dopants in silicon, for which electrical read-out based on a single-electron transistor is used[1], and nitrogen–vacancy centres in diamond, for which optical read-out is used[4–6]. A spin read-out fidelity of about 90 per cent has been demonstrated with both electrical read-out[1] and optical read-out[10,11]; however, the thermal limitations of the former and the poor photon collection efficiency of the latter make it difficult to achieve the higher fidelities required for quantum information applications. Here we demonstrate a hybrid approach in which optical excitation is used to change the charge state (conditional on its spin state) of an erbium defect centre in a silicon-based single-electron transistor, and this change is then detected electrically. The high spectral resolution of the optical frequency-addressing step overcomes the thermal broadening limitation of the previous electrical read-out scheme, and the charge-sensing step avoids the difficulties of efficient photon collection. This approach could lead to new architectures for quantum information processing devices and could drastically increase the range of defect centres that can be exploited. Furthermore, the efficient electrical detection of the optical excitation of single sites in silicon represents a significant step towards developing interconnects between optical-based quantum computing and silicon technologies.

The potential for hybrid optical–electrical single-spin read-out was recently established, and a long nuclear spin coherence time demonstrated for an ensemble of P ions in highly purified ^{28}Si (ref. 12). The spin ensemble was read out by detecting the photocurrent generated when the 1,078-nm excitonic transition associated with the P ions was excited. In the present work, we demonstrate single-site detection by electrically detecting the optical excitation of the $^4I_{15/2}$–$^4I_{13/2}$ transition of single Er ions implanted into silicon, resolving both electronic Zeeman and hyperfine structure. The efficient read-out required for single-site detection is achieved by measuring the photo-induced change in the site's charge state using a single-electron transistor (SET), rather than by detecting the associated photocurrent.

The large magnetic moment of erbium's electronic ground state, the $I = 7/2$ nuclear spin of the ^{167}Er isotope and the coincidence between the $^4I_{15/2}$–$^4I_{13/2}$ transition and the 1.5-μm transmission window of silica optical fibres make Er centres appealing for quantum information applications[13–15]. The few existing studies in samples with a high concentration of Er have shown nuclear spin relaxation times of 0.1 s (ref. 16) and electron spin dephasing times of 100 μs (ref. 15). Single-site detection grants access to low Er densities, where we expect drastically enhanced coherence times in analogy to the recent development for P in Si[12]. The low emission rate from optically excited rare-earth ions such as Er^{3+} makes pure optical detection of single sites challenging[13,14]. Recently, however, the optical detection of a single rare-earth ion was demonstrated in a Pr-doped YAG nanocrystal[17]. The technique used involved the two-step excitation of the ion to a high-lying 5*d* electron state and detection of the resultant emission[17]. The experiment was conducted at room temperature and resulted in low detection efficiency and low frequency resolution, making state read-out infeasible.

The $^4I_{15/2}$–$^4I_{13/2}$ transition is between states within the inner 4*f* electron shell of the Er^{3+} ion, which is well shielded from the surrounding lattice by filled outer shells, resulting in narrow spectral linewidths and the potential for high-resolution frequency addressing. At liquid helium temperatures, homogeneous linewidths as narrow as 50 Hz have been observed for the transition in Er^{3+}:Y_2SiO_5 (ref. 18). Until now, there have not been any sub-inhomogeneous-linewidth studies conducted on optical transition in Er centres in silicon. The observed lifetime of emission of 2 ms from the $^4I_{13/2}$ state for Er^{3+} ions in silicon implies a minimum linewidth of 150 Hz (ref. 19).

The resonant photoionization of individual Er^{3+} ions is studied in an Er-implanted SET (Fig. 1a), which works as a charge sensor. The $^4I_{15/2}$–$^4I_{13/2}$ transition of an Er^{3+} ion has a relatively high probability, when a stimulating laser is tuned to its resonant wavelength, and the Er^{3+} ion could be further ionized owing to a two-photon process or an Auger process. The charge displacement induced by an ionization event simultaneously leads to a change in the tunnelling current of the SET. To get a high sensitivity, the SET is biased close to the degeneracy point between two charge states, that is, at the edge of one Coulomb peak (Fig. 1b). Accordingly, the transconductance is large, and a small charge displacement in the sensitive region will lead to a significant change in the tunnelling current[20,21]. The photoionization of individual Er^{3+} ions leads to a significant change in tunnelling current (Fig. 1b). The $^4I_{15/2}$–$^4I_{13/2}$ transition of each Er^{3+} ion has a specific resonant photon energy, by which individual Er^{3+} ions can be distinguished. When the laser is tuned to a non-resonant wavelength, the tunnelling current mainly stays at the background level (Fig. 1c). In contrast, when the laser is tuned to a resonant wavelength of an Er^{3+} ion, the photoionization of the Er^{3+} ion leads to a rise in the tunnelling current. The current then returns to the lower level owing to its neutralization, resulting in a two-level current–time trace (Fig. 1d), which suggests that only one single Er^{3+} ion is ionized (Methods). Figure 1e shows a photoionization spectrum of a single Er^{3+} ion. Current–time traces are recorded at a series of photon energies, and then the histogram showing the distribution of binned current in time is plotted as a function of the photon energy detuning. The colour in Fig. 1e represents the time (Σt_{bin}) during which the current stays within one bin, and a 0.02-nA bin size is used throughout the analysis.

As shown in Fig. 1a, the SET has a Si channel that passes under the gate. The SET is biased below the threshold voltage, so that the current tunnels through the corner regions of the Si channel[22]. Consequently, the charge sensor is more sensitive to the Er^{3+} ions that are closer to the corner regions in the channel, and different Er^{3+} ions have different capacitive couplings leading to different detection sensitivities. The change in current (Fig. 1b–d) accords with the loss of an electron, indicating that it is due to the ionization of the Er centre, whereas

[1]Centre of Excellence for Quantum Computation and Communication Technology, School of Physics, University of New South Wales, Sydney, New South Wales 2052, Australia. [2]Centre of Excellence for Quantum Computation and Communication Technology, RSPE, Australian National University, Canberra, Australian Capital Territory 0200, Australia. [3]Centre of Excellence for Quantum Computation and Communication Technology, School of Physics, University of Melbourne, Melbourne, Victoria 3010, Australia.

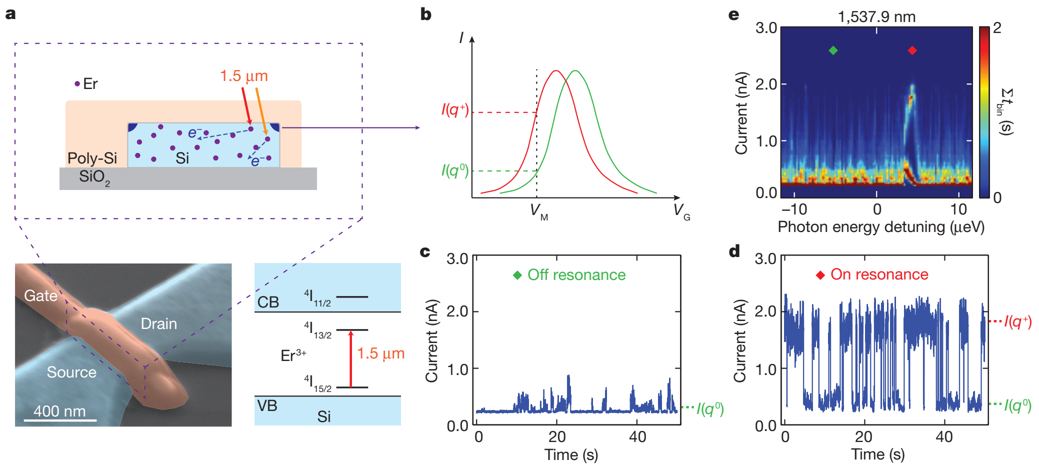


Figure 1 | Photoionization spectroscopy of an individual Er^{3+} ion. **a**, Coloured scanning electron micrograph of a typical SET device used in this study and a band structure of Er^{3+} ions in silicon. CB, conduction band; VB, valence band. Top, schematic cross-section of the SET showing the optical addressing of individual Er^{3+} ions. **b**, The SET charge-sensing scheme. The loss of an electron due to photoionization induces a transient shift of the current (I)/gate voltage (V_G) curve towards lower gate voltages, causing a change in current from $I(q^0)$ to $I(q^+)$. **c**, **d**, The current–time traces recorded for a fixed gate voltage (V_M) under non-resonant (**c**) and resonant (**d**) illumination. **e**, The histogram of current–time traces as a function of the photon energy detuning. The photon energy of the illumination is detuned with respect to the centred wavelength of 1,537.9 nm.

the gain of an electron will lead to a shift opposite to that in Fig. 1b, that is, to higher gate voltages. The small fluctuations in current, which we attribute to the trap states (defects that can capture electrons) in the gate dielectric or the oxide layer with weak capacitive coupling[23], can be suppressed by properly annealing the devices before fabrication. The read-out efficiency is mainly limited by the efficiency of the excitation from the $^4I_{13/2}$ excited state into the conduction band, which can be increased to close to 100% by increasing the intensity of the light used to drive this final ionization step. We observe resonances, using photoionization spectroscopy, mostly between 1,535 and 1,539 nm, which is consistent with the $^4I_{15/2}$–$^4I_{13/2}$ transition of Er^{3+} ions in silicon[13,14].

We next study the Zeeman effect of individual Er^{3+} ions, because the Zeeman effect is an essential tool to determine the site symmetry of Er centres. Er ions in Si tend to have a valence of 3+, characteristic of the Si lattice, and so the 4*f* electrons of Er^{3+} ions have the ground state $^4I_{15/2}$ and the first excited state $^4I_{13/2}$ (ref. 13). The degeneracy is lifted by the crystal field, such that each state splits into several levels depending on the symmetry of the Er centre[13]. The transition between the lowest level of $^4I_{15/2}$ and the lowest level of $^4I_{13/2}$ is responsible for the strong emission band around 1.54 μm, and the Zeeman splitting of those two levels in the case of double degeneracy is shown in Fig. 2a. The doublet states can be described by an effective spin of $S = 1/2$, and the Zeeman interaction has the form $H = \beta_e BgS$, where β_e is the electronic Bohr magneton, B is the magnetic field, g is the g-factor matrix[24] and S is the spin vector. The Zeeman splitting energies of the higher- and lower-energy doublets are $\Delta E_e^H = \beta_e g_H B$ and $\Delta E_e^L = \beta_e g_L B$, respectively. As shown in Fig. 2a, two types of optical transition can be excited between these doublets. The photon energy difference between two transitions of the same type can be described by $\Delta E_{photon} = \beta_e \Delta g B$, where Δg, the g-factor difference, is $|g_H + g_L|$ for $\Delta M_S = \pm 1$ transitions and $|g_H - g_L|$ for $\Delta M_S = 0$ transitions. In this study, we measured the Zeeman splitting of four spectrally isolated Er resonances, and observed g-factor differences ranging from 1.6 to 10.8.

Figure 2b, c shows the Zeeman splitting of Er^{3+} ions. Current–time traces are recorded at a series of photon energies and magnetic fields; each pixel in Fig. 2b, c represents one trace. When an Er^{3+} ion is ionized, the current will exceed a certain threshold, which is determined by the background current fluctuation under non-resonant illumination. For each current–time trace, the time (t_{up}) during which the current exceeds the threshold is integrated and gives the values (Σt_{up}) plotted in Fig. 2b, c. As shown in Fig. 2b, the resonance occurs at a photon energy detuning of 1 μeV away from 1,535.8 nm at zero magnetic field, and starts to split into two diagonal arms as the magnetic field increases. It is due to the Zeeman effect of one individual Er^{3+} ion, with $\Delta g \approx 4.8$. Similarly, the Zeeman splitting of the resonance around 1,538.0 nm is studied as shown in Fig. 2c, where the darkest blue rectangular regions were not scanned. There seem to be two resonances with similar resonant wavelengths and the same g-factor difference ($\Delta g \approx 3.3$), but with different signal intensities. This could be due to two individual Er^{3+} ions with the same site symmetry but with different capacitive couplings. Furthermore, the Zeeman splitting of the resonance around 1,538.0 nm shows polarization dependence. As shown in

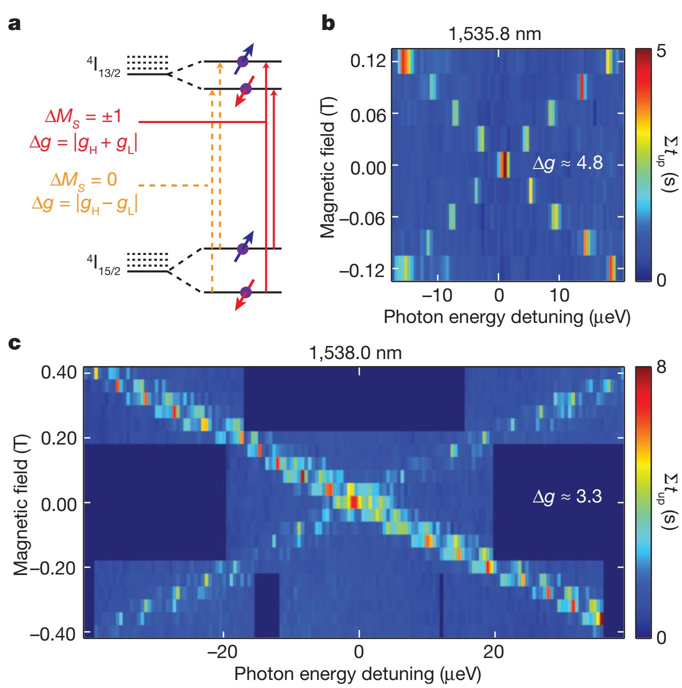


Figure 2 | The Zeeman effect of individual Er^{3+} ions. **a**, Schematic diagrams showing the Zeeman splitting and optical transitions of Er^{3+} ions in silicon. The splitting of the $^4I_{13/2}$ and $^4I_{15/2}$ states depends on the site symmetry of the Er centre. **b**, The Zeeman splitting scan of the Er resonance with a centred wavelength of 1,535.8 nm. Each pixel represents a current–time trace recorded for 50 s. **c**, The Zeeman splitting scan of the Er resonance with a centred wavelength of 1,538.0 nm.

Fig. 2c, the diagonal arm is weaker than the anti-diagonal one. By modifying the polarization of the light entering the cryostat, we tuned the diagonal arm to be stronger than the anti-diagonal one. The site symmetry of individual Er centres can be determined from the polarization dependence and a rotating magnetic field measurement. Spin-selective excitation even for degenerate spin states can be achieved by maximizing the contrast between two Zeeman arms (Fig. 2c), which allows spin read-out without a magnetic field.

The hyperfine structure is of great interest because the nuclear spin has long coherence times useful for quantum information storage[12,25,26]. In addition, this structure is an effective means of distinguishing between different ions as well as other defects. Erbium has six stable isotopes, among which only ^{167}Er has a non-zero nuclear spin, of $I = 7/2$, leading to eight nuclear spin states. At high magnetic field, the hyperfine interaction can be treated as a perturbation of the Zeeman effect[27], and so each electron spin state will split into eight sublevels owing to the hyperfine interaction (Fig. 3a). We denote the total splitting energies of these eight sublevels of $^4I_{15/2}$ and $^4I_{13/2}$ as ΔE_N^L and ΔE_N^H, respectively. At low magnetic field, the hyperfine interaction is comparable to the Zeeman effect, so the sublevels will mix[28].

To investigate the hyperfine structure of $^{167}Er^{3+}$ ions, we implanted ^{167}Er ions and ^{168}Er (with zero nuclear spin) control ions. We first study the photoionization spectrum of an Er^{3+} control ion with zero nuclear spin. The integrated time (Σt_{up}) is plotted as a function of the photon energy detuning, as indicated by the blue dashed line in Fig. 3b. The same spectral asymmetry as that in Fig. 1e is observed. We attribute the asymmetry to the correlation between the Stark shift of the $^4I_{15/2}$–$^4I_{13/2}$ resonance and a broadening of the Coulomb peak, both of which are sensitive to fluctuating electric fields in the channel. The fluctuating field is attributed to the laser excitation of trap states in or near the channel. Because we directly observe the effect on the Coulomb peak, it is possible to remove part of this broadening of the peak. After applying this correction (Methods), a minimum spectral full-width at half-maximum of 50 neV is observed as indicated by the red solid line in Fig. 3c. To significantly reduce the linewidth further, it will be necessary to reduce the density of trap states. As well as undergoing electric-field-induced shifts, the line is expected to be broadened through magnetic interactions with both ^{29}Si, by up to tens of nanoelectronvolts, and paramagnetic centres in the device. It is expected from analogy with observations in Er^{3+}:Y_2SiO_5 (ref. 18) that applying a large magnetic field will suppress broadening by this mechanism.

We now show the hyperfine structure of one $^{167}Er^{3+}$ ion. The photoionization spectrum recorded at high magnetic field ($B = 0.14$ T) (Fig. 3c) reveals eight resonant peaks with a photon energy difference of about 0.2 μeV between each successive pair. The high spectral resolution allows nuclear spin read-out with potential for single-shot read-out and manipulation of the nuclear spin states. Because the addressability does not rely on a specific magnetic field, the photoionization spectra are measured at a series of magnetic fields (Fig. 3d). The Zeeman shift is subtracted to show the evolution of the hyperfine interaction. At high magnetic field, eight significant peaks are observed in total (for $0.08\,T \leq B \leq 0.14\,T$), because here the hyperfine interaction can be treated as a perturbation of the Zeeman effect. At low magnetic field, multiple resonances appear (for $-0.04\,T \leq B \leq 0.06\,T$), revealing the mixing of the hyperfine sublevels, because here the hyperfine interaction is comparable to the Zeeman effect.

The eight significant peaks, representing the eight different nuclear spin states of ^{167}Er, demonstrate that the resonances are due to the $^{167}Er^{3+}$ ion rather than to other ions or defects. The peaks (Fig. 3c) correspond to the allowed transitions ($\Delta M_I = 0$) preserving the nuclear spin states, but the question remains of whether they are due to the $\Delta M_S = 0$ transitions or the $\Delta M_S = \pm 1$ transitions. As shown in Fig. 3a, we attribute them to the $\Delta M_S = 0$ transitions for two reasons. First, the energy difference between the two most distant hyperfine peaks is only about 1.7 μeV (Fig. 3c), which is much smaller than the typical splitting energy of the $\Delta M_S = \pm 1$ transitions of Er^{3+} ions. Electron paramagnetic resonance measurements on $^{167}Er^{3+}$ ions in crystals show a splitting energy ($2\Delta E_N^L$) of about 30 μeV (refs 15, 29), which corresponds to the $\Delta M_S = \pm 1$ transitions. Second, a ninth peak appears beyond the region between the two most distant peaks (at a photon energy detuning of −1.1 μeV in Fig. 3c). This peak is recognizable but is much weaker than the other eight peaks, a weakness that we attribute to a forbidden transition. The energy of the forbidden transitions ($\Delta M_I = \pm 1$) of Er^{3+} ions can lie outside the region between the two most distant peaks of the allowed transitions, but only in the case of the $\Delta M_S = 0$ transitions. Consequently, the eight significant peaks are attributed to the $\Delta M_S = 0$ transitions, and the splitting energy is expressed as $|\Delta E_N^H - \Delta E_N^L| = 1.7$ μeV.

Hybrid optical–electrical access to single spins of individual ions in a nanotransistor has been demonstrated, and is applicable to other defects in solids. Specifically, with an Er-implanted SET, photoionization spectroscopy allows real-time observation of single optical excitation events, avoiding the bottleneck of photon collection. Furthermore, high-resolution optical frequency addressing circumvents the limitations due to thermal broadening in earlier electrical detection of impurity spins[1]. Our findings open the way to the optical addressing and manipulation of the electron and nuclear spin states of individual defects in a solid, other than nitrogen–vacancy centres in diamond. In addition, this hybrid optical–electrical technique advances the microstructural study of ions in a semiconductor to the single-site level, including microscopic aspects, electrical and optical activity, and so on.

An approach that combines dopant ions (for example Er and P) with quantum optical control and semiconductor fabrication technologies represents an attractive platform with which to realize a scalable quantum computation and communication architecture. Such a system could consist of individual ions inside a ring cavity coupled to

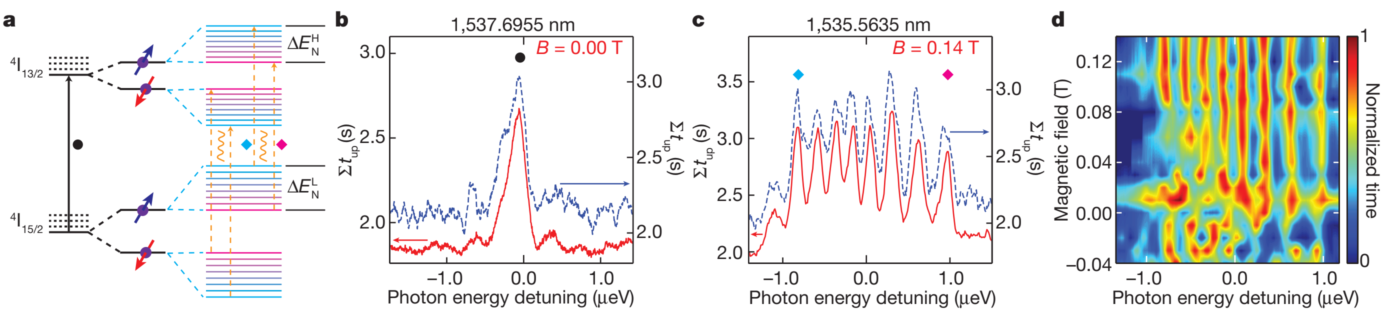


Figure 3 | The hyperfine structure of an individual Er^{3+} ion. **a**, Schematic diagrams showing the hyperfine splitting and the $\Delta M_S = 0$ transitions (orange dashed line) of $^{167}Er^{3+}$ ions at high magnetic field, and, for comparison, the optical transition of Er^{3+} ions with zero nuclear spin at zero magnetic field (black solid line). **b**, The photoionization spectrum of an Er^{3+} control ion with zero nuclear spin. The red solid and blue dashed lines respectively represent the data with and without removing the broadening. **c**, The photoionization spectrum of a single $^{167}Er^{3+}$ ion. The eight significant peaks correspond to optical transitions of the eight nuclear spin states of ^{167}Er ($I = 7/2$). **d**, The contour plot of the photoionization spectra of the $^{167}Er^{3+}$ ion, showing evolution of the hyperfine interaction.

each other via photons, with nearby charge-sensing devices used to read out the spin states of individual ions and to control the coupling between ions by Stark tuning. The ring cavities could be connected by optical waveguides, which enable quantum information transfer between individual ions in different cavities. Here we have demonstrated the first step towards such a system, that is, optical addressing of individual ions, and further improvement can be made by reducing the observed linewidth as discussed above. However, there are essential issues to be addressed in the future, such as electron and nuclear spin coherence times of Er and P ions, the influence of photoionization on nuclear spin coherence and spin–photon entanglement.

Note added in proof: During the production of the manuscript we became aware of new all-optical single-spin detection of defect centres (ref. 31 and F. Jelezko, personal communication), which further illustrates the attractiveness of a hybrid electrical–optical approach.

METHODS SUMMARY

The devices were fabricated at IMEC as previously described[30]. After complete device fabrication, an Er:O co-implantation (dose ratio, 1:6) was performed with implantation energies of 400 keV (Er) and 55 keV (O). There should be approximately 30–40 Er ions in the sensitive region of one Coulomb peak. Under the erbium implantation conditions we used, the beam is estimated to have been composed of 70–80% ^{168}Er and 20–30% ^{167}Er. The devices were then annealed at 700 °C in N_2 for 10 min to remove implantation damage and to initiate the formation of Er centres. All the measurements were carried out in a liquid helium cryostat at 4.2 K. The laser beam, with an optical power of 4–5 mW, passed through a single-mode fibre and was incident on the sample with a beam diameter of about 1 mm. In the initial phase of the experiments (Figs 1 and 2), a commercial tunable laser with an external cavity was used. To maintain a high precision, we set one centred wavelength using a motor-actuator and swept the wavelength about the centred wavelength using a piezo-actuator. In the high-resolution experiments (Fig. 3), the wavelength of another laser was stabilized to about 0.01 pm, and a wavelength meter was used to compensate the thermal drift.

Full Methods and any associated references are available in the online version of the paper.

Received 12 December 2012; accepted 18 March 2013.

1. Morello, A. *et al.* Single-shot readout of an electron spin in silicon. *Nature* **467,** 687–691 (2010).
2. Fuechsle, M. *et al.* A single-atom transistor. *Nature Nanotechnol.* **7,** 242–246 (2012).
3. Pla, J. J. *et al.* A single-atom electron spin qubit in silicon. *Nature* **489,** 541–545 (2012).
4. Gaebel, T. *et al.* Room-temperature coherent coupling of single spins in diamond. *Nature Phys.* **2,** 408–413 (2006).
5. Jiang, L. *et al.* Repetitive readout of a single electronic spin via quantum logic with nuclear spin ancillae. *Science* **326,** 267–272 (2009).
6. Togan, E. *et al.* Quantum entanglement between an optical photon and a solid-state spin qubit. *Nature* **466,** 730–734 (2010).
7. Maze, J. R. *et al.* Nanoscale magnetic sensing with an individual electronic spin in diamond. *Nature* **455,** 644–647 (2008).
8. Morton, J. J. L., McCamey, D. R., Eriksson, M. A. & Lyon, S. A. Embracing the quantum limit in silicon computing. *Nature* **479,** 345–353 (2011).
9. Zwanenburg, F. A. *et al.* Silicon quantum electronics. Preprint at http://arxiv.org/abs/1206.5202 (2012).
10. Robledo, L. *et al.* High-fidelity projective read-out of a solid-state spin quantum register. *Nature* **477,** 574–578 (2011).
11. Neumann, P. *et al.* Single-shot readout of a single nuclear spin. *Science* **329,** 542–544 (2010).
12. Steger, M. *et al.* Quantum information storage for over 180 s using donor spins in a ^{28}Si "semiconductor vacuum". *Science* **336,** 1280–1283 (2012).
13. Kenyon, A. J. Erbium in silicon. *Semicond. Sci. Technol.* **20,** R65 (2005).
14. Vinh, N. Q., Ha, N. N. & Gregorkiewicz, T. Photonic properties of Er-doped crystalline silicon. *Proc. IEEE* **97,** 1269–1283 (2009).
15. Bertaina, S. *et al.* Rare-earth solid-state qubits. *Nature Nanotechnol.* **2,** 39–42 (2007).
16. Baldit, E. *et al.* Identification of Λ-like systems in Er^{3+}:Y_2SiO_5 and observation of electromagnetically induced transparency. *Phys. Rev. B* **81,** 144303 (2010).
17. Kolesov, R. *et al.* Optical detection of a single rare-earth ion in a crystal. *Nature Commun.* **3,** 1029 (2012).
18. Sun, Y., Thiel, C. W., Cone, R. L., Equall, R. W. & Hutcheson, R. L. Recent progress in developing new rare earth materials for hole burning and coherent transient applications. *J. Lumin.* **98,** 281–287 (2002).
19. Priolo, F., Franz, G., Coffa, S. & Carnera, A. Excitation and nonradiative deexcitation processes of Er^{3+} in crystalline Si. *Phys. Rev. B* **57,** 4443 (1998).
20. Pioda, A. *et al.* Single-shot detection of electrons generated by individual photons in a tunable lateral quantum dot. *Phys. Rev. Lett.* **106,** 146804 (2011).
21. Hanson, R., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Spins in few-electron quantum dots. *Rev. Mod. Phys.* **79,** 1217–1265 (2007).
22. Sellier, H. *et al.* Subthreshold channels at the edges of nanoscale triple-gate silicon transistors. *Appl. Phys. Lett.* **90,** 073502 (2007).
23. Tettamanzi, G. C. *et al.* Interface trap density metrology of state-of-the-art undoped Si n-FinFETs. *IEEE Electron Device Lett.* **32,** 440–442 (2011).
24. Guillot-Nol, O. *et al.* Hyperfine interaction of Er^{3+} ions in Y_2SiO_5: an electron paramagnetic resonance spectroscopy study. *Phys. Rev. B* **74,** 214409 (2006).
25. Hedges, M. P., Longdell, J. J., Li, Y. & Sellars, M. J. Efficient quantum memory for light. *Nature* **465,** 1052–1056 (2010).
26. Simmons, S. *et al.* Entanglement in a solid-state spin ensemble. *Nature* **470,** 69–72 (2011).
27. Smith, K. F. & Unsworth, P. J. The hyperfine structure of ^{167}Er and magnetic moments of ^{143,145}Nd and ^{167}Er by atomic beam triple magnetic resonance. *Proc. Phys. Soc.* **86,** 1249 (1965).
28. McAuslan, D. L., Bartholomew, J. G., Sellars, M. J. & Longdell, J. J. Reducing decoherence in optical and spin transitions in rare-earth-metal-ion–doped materials. *Phys. Rev. A* **85,** 032339 (2012).
29. Yang, S. *et al.* Electron paramagnetic resonance of Er^{3+} ions in aluminum nitride. *J. Appl. Phys.* **105,** 023714 (2009).
30. Lansbergen, G. P. *et al.* Gate-induced quantum-confinement transition of a single dopant atom in a silicon FinFET. *Nature Phys.* **4,** 656–661 (2008).
31. Sleiter, D. J. *et al.* Optical pumping of a single electron spin bound to a fluorine donor in a ZnSe nanostructure. *Nano Lett.*. **13**(1), 116–120 (2013).

Acknowledgements We thank R. Ahlefeldt, J. Bartholomew, R. Elliman, N. Manson and A. Morello for discussions. We also thank M. Hedges and T. Lucas for their help in the initial phase of the experiments. The devices were fabricated by N. Collaert and S. Biesemans. This work was financially supported by the ARC Centre of Excellence for Quantum Computation and Communication Technology (CE110001027) and the Future Fellowships (FT100100589 and FT110100919).

Author Contributions N.S. and J.C.M. designed and performed the implantation. C.Y., M.J.S. and S.R. designed and conducted the experiments. C.Y., M.R. and G.G.d.B. carried out the experiments. All the authors contributed to analysing the results and writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.R. (s.rogge@unsw.edu.au).

METHODS

Details of the devices. The devices used in this study are n-p-n field-effect transistors fabricated at IMEC as previously described[30]. Each device has a p-type silicon channel passing under a polycrystalline silicon gate separated by a gate dielectric. The p-type channel has a boron doping of $10^{18}\,\mathrm{cm}^{-3}$. After complete device fabrication, an Er:O co-implantation is performed with respective implantation energies of 400 and 55 keV and ion fluences of 4×10^{12} and $3\times10^{13}\,\mathrm{cm}^{-2}$. This leads to an Er:O dose ratio of about 1:6 in the channel region. Under the erbium implantation conditions we use, the beam is estimated to have been composed of 70–80% ^{168}Er and 20–30% ^{167}Er. The presence of both oxygen impurities[13] and boron impurities[32] is known to enhance the luminescence of the Er^{3+} ions in silicon. The 700 °C post-implantation anneal is within the thermal processing window for Er centre activation in silicon[32].

In the experiments, the device is biased below the threshold voltage, and only the corner regions of the silicon channel go into inversion[22]. A peak of the I–V_G curve is due to the Coulomb blockade in one of the two corner regions, where the current flows. The sensitive region is defined as the region in which one elementary charge change can be detected, taking the current noise and the transconductance of the Coulomb peak into account. The sensitive region of one Coulomb peak is estimated to be the corresponding channel-corner region, which, for the device shown in Fig. 1a, has dimensions of 100 nm × 50 nm × 20 nm (length × width × height). Simulations of the ion implantation based on SRIM[33] show that there should be approximately 30–40 Er ions in the sensitive region of one Coulomb peak.

Experimental details and data analysis. All the measurements are carried out in a liquid helium cryostat at 4.2 K. The laser beam, with an optical power of 4–5 mW, passes through a single-mode fibre and is incident on the sample with a beam diameter of about 1 mm. In the initial phase of the experiments (Figs 1 and 2), a commercial tunable laser with an external cavity is used. To maintain a high precision, we set one centred wavelength using a motor-actuator, and sweep the wavelength about the centred wavelength using a piezo-actuator. The current–time traces in Fig. 1c and Fig. 1d are recorded at two different photon energies and are consistent with the photoionization spectrum as indicated by the green and red diamonds in Fig. 1e, respectively. For instance, the current mainly stays at the background level (0.4 nA) at a photon energy detuning of −5 μeV, but jumps between two levels (1.8 and 0.4 nA) at a photon energy detuning of 4 μeV. It is worth noting that the two-level trace (Fig. 1d) suggests that only a single Er^{3+} ion is ionized. Multiple ions with different capacitive couplings will lead to current–time traces with more than two levels, and two ions with the same capacitive coupling will lead to a current–time trace with three levels once they are simultaneously ionized. We attribute the charge displacement to the ionization of an Er^{3+} ion rather than the charge fluctuations of the trap states, because all the Er^{3+} ions that we observe contribute to a shift of the Coulomb peak towards lower gate voltages. In the high-resolution experiments (Fig. 3), the wavelength of another laser is stabilized to about 0.01 pm, and a wavelength meter is used to compensate the thermal drift. The asymmetry as well as part of the broadening of the resonant peak is removed by adding an photon energy offset to the data. We then integrate the time during which the current exceeds the threshold to obtain the values plotted as the red solid line in Fig. 3b, c. For comparison, with the broadening removed the red solid line shows smaller widths and less noise than does the blue dashed line without the broadening removed. Nevertheless, the resonances in the latter are still clearly visible (Fig. 3b, c).

32. Michel, J. *et al.* Impurity enhancement of the 1.54 μm Er^{3+} luminescence in silicon. *J. Appl. Phys.* **70,** 2672–2678 (1991).
33. Ziegler, J. F., Ziegler, M. & Biersack, J. SRIM – The stopping and range of ions in matter (2010). *Nucl. Instrum. Methods Phys. Res. B* **268,** 1818–1823 (2010).

LETTER

doi:10.1038/nature12083

Digital cameras with designs inspired by the arthropod eye

Young Min Song[1]*, Yizhu Xie[1]*, Viktor Malyarchuk[1]*, Jianliang Xiao[2]*, Inhwa Jung[3], Ki-Joong Choi[4], Zhuangjian Liu[5], Hyunsung Park[6], Chaofeng Lu[7,8], Rak-Hwan Kim[1], Rui Li[8,9], Kenneth B. Crozier[6], Yonggang Huang[8] & John A. Rogers[1,4]

In arthropods, evolution has created a remarkably sophisticated class of imaging systems, with a wide-angle field of view, low aberrations, high acuity to motion and an infinite depth of field[1–3]. A challenge in building digital cameras with the hemispherical, compound apposition layouts of arthropod eyes is that essential design requirements cannot be met with existing planar sensor technologies or conventional optics. Here we present materials, mechanics and integration schemes that afford scalable pathways to working, arthropod-inspired cameras with nearly full hemispherical shapes (about 160 degrees). Their surfaces are densely populated by imaging elements (artificial ommatidia), which are comparable in number (180) to those of the eyes of fire ants (*Solenopsis fugax*) and bark beetles[4,5] (*Hylastes nigrinus*). The devices combine elastomeric compound optical elements with deformable arrays of thin silicon photodetectors into integrated sheets that can be elastically transformed from the planar geometries in which they are fabricated to hemispherical shapes for integration into apposition cameras. Our imaging results and quantitative ray-tracing-based simulations illustrate key features of operation. These general strategies seem to be applicable to other compound eye devices, such as those inspired by moths and lacewings[6,7] (refracting superposition eyes), lobster and shrimp[8] (reflecting superposition eyes), and houseflies[9] (neural superposition eyes).

Improved understanding of light-sensing organs in biology[1,10–12] creates opportunities for the development of cameras that adopt similar engineering principles, to provide operational characteristics beyond those available with existing technologies[13–19]. The compound eyes of arthropods are particularly notable for their exceptionally wide fields of view, high sensitivity to motion and infinite depth of field[1–3]. Analogous man-made cameras with these characteristics have been of long-standing interest, owing to their potential for use in surveillance devices, tools for endoscopy and other demanding applications. Previous work demonstrates devices that incorporate compound lens systems, but only in planar geometries or in large-scale, handmade curved replicas[20–24]. Constraints intrinsic to such approaches prevent the realization of cameras with the key features present in arthropod eyes: full hemispherical shapes in compact, monolithic forms, with scalability in size, number and configuration of the light-sensing elements (ommatidia).

Here we present a complete set of materials, design layouts and integration schemes for digital cameras that mimic hemispherical apposition compound eyes found in biology. Certain of the concepts extend recent advances in stretchable electronics[25] and hemispherical photodetector arrays[13–18], in overall strategies that provide previously unachievable options in design. Systematic experimental and theoretical studies of the mechanical and optical properties of working devices reveal the essential aspects of fabrication and operation.

Figure 1a presents schematic illustrations of the two main subsystems and methods for their assembly into working hemispherical apposition cameras. The first subsystem provides optical imaging function and defines the overall mechanics; it is a moulded piece of the elastomer poly(dimethylsiloxane) (PDMS, with index of refraction $n \approx 1.43$) that consists of an array of 16×16 convex microlenses (with radius of curvature of each microlens $r \approx 400\ \mu m$) over a square area of $14.72\ mm \times 14.72\ mm$, as shown in Supplementary Fig. 1. Of the 256 microlenses, 180 form working components of the camera, each on a matching cylindrical supporting post (of height $h \approx 400\ \mu m$) connected to a base membrane (of thickness $t \approx 550\ \mu m$).

The second subsystem enables photodetection and electrical readout; it consists of a matching array of thin, silicon photodiodes (active areas $d^2 \approx 160\ \mu m \times 160\ \mu m$) and blocking diodes in an open mesh configuration with capability for matrix addressing. Narrow filamentary serpentine traces of metal (Cr/Au) encapsulated by polyimide serve as electrical and mechanical interconnects. Aligned bonding of these two subsystems places each photodiode at the focal position of a corresponding microlens (Fig. 1b), to yield an integrated imaging system. A key feature enabled by the constituent materials and layouts is a fully isotropic elastic mechanical response to large strain deformation, in any direction. In consequence, hydraulic actuation can deterministically transform the planar layout in which these separate subsystems are constructed and bonded together, into a full hemispherical shape (Fig. 1c), with precise engineering control (radius of curvature of the hemisphere $R \approx 6.96$ mm) and without any change in optical alignment or adverse effect on electrical or optical performance (see Supplementary Figs 2 and 3 for details).

A complete apposition camera (Fig. 1d) consists of this type of imager, combined with a top perforated sheet and a bottom bulk support, both made of black silicone to eliminate stray light, bonded to the outer and inner surfaces, respectively (Fig. 1e and Supplementary Fig. 8). A thin film insert with metallized contact pads connects to a printed circuit board as an interface to external control electronics.

By analogy to imaging organs in arthropods[12], each microlens and supporting post corresponds to a corneal lens and crystalline cone, respectively; each photodiode is a rhabdom; the black elastomer serves as the screening pigment. A collected set of each of these elements represents an ommatidium.

The dimensions and the mechanical properties of the imaging system are critically important for proper operation. The acceptance angle ($\Delta\varphi$) and the inter-ommatidial angle ($\Delta\Phi$) define the nature of image formation[1,26] (Fig. 1a and Supplementary Fig. 4). Each microlens focuses light incident on it within a cone defined by $\Delta\varphi$. An individual ommatidium samples an angular object space determined by $\Delta\Phi$. For the layouts of Fig. 1a, optical simulation suggests a total field of view of

[1]Department of Materials Science and Engineering, Beckman Institute for Advanced Science and Technology, and Frederick Seitz Materials Research Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [2]Department of Mechanical Engineering, University of Colorado at Boulder, Boulder, Colorado 80309, USA. [3]Department of Mechanical Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 446-701, South Korea. [4]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [5]Institute of High Performance Computing, A*star, 1 Fusionopolis Way, #16-16 Connexis 138632, Singapore. [6]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. [7]Department of Civil Engineering and Soft Matter Research Center, Zhejiang University, Hangzhou 310058, China. [8]Department of Mechanical Engineering, Department of Civil and Environmental Engineering, Northwestern University, Evanston, Illinois 60208, USA. [9]State Key Laboratory of Structural Analysis for Industrial Equipment, Department of Engineering Mechanics, Dalian University of Technology, Dalian 116024, China.
*These authors contributed equally to this work.

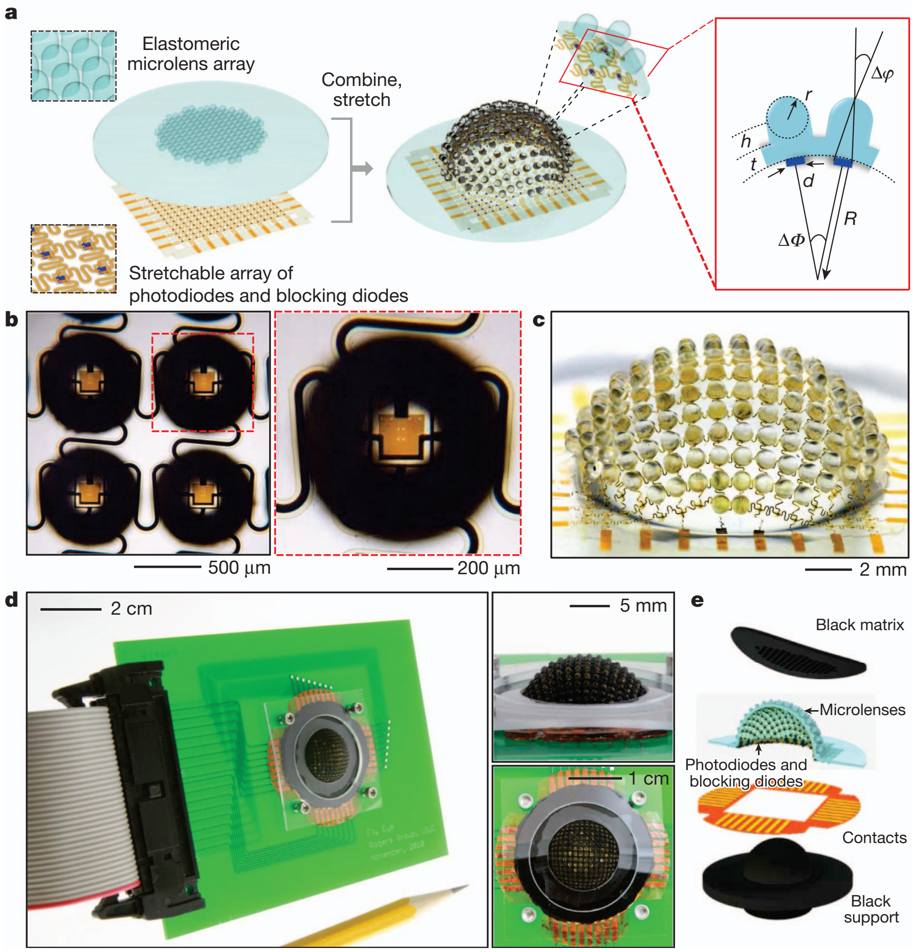


Figure 1 | Schematic illustrations and images of components and integration schemes for a digital camera that takes the form of a hemispherical, apposition compound eye. **a**, Illustrations of an array of elastomeric microlenses and supporting posts joined by a base membrane (above) and a corresponding collection of silicon photodiodes and blocking diodes interconnected by filamentary serpentine wires and configured for matrix addressing (below). On the left, these components are shown in their as-fabricated, planar geometries; the upper and lower insets provide magnified views of four adjacent unit cells (that is, artificial ommatidia). Bonding these two elements and elastically deforming them ('combine, stretch') into a hemispherical shape creates the digital imaging component of an apposition compound eye camera (centre). An exploded view of four adjacent unit cells appears in the centre inset, with a cross-sectional illustration (on the right) that highlights key parameters: the acceptance angle ($\Delta\varphi$) for each ommatidium, the inter-ommatidial angle ($\Delta\Phi$), the radius of curvature of the entire device (R) and of an individual microlens (r), the height of a cylindrical supporting post (h), the thickness of the base membrane (t), and the diameter of the active area of a photodiode (d). **b**, Optical micrograph of four adjacent ommatidia in planar format (left), with magnified view (right). **c**, Image of a representative system after hemispherical deformation. **d**, Photograph of a completed camera mounted on a printed circuit board as an interface to external control electronics (left), with close-ups in the insets (upper inset shows tilted view; lower inset shows top view). **e**, Exploded-view illustration of the components of this system: perforated sheet of black silicone (black matrix), hemispherical array of microlenses and photodiodes/blocking diodes, thin-film contacts for external interconnects, and hemispherical supporting substrate of black silicone.

about 160° when $\Delta\varphi = 9.7°$ and $\Delta\Phi = 11.0°$, without overlapping fields in adjacent ommatidia (that is, $\Delta\varphi < \Delta\Phi$; see Supplementary Figs 5 and 6 for details). The key dimensions of the optical subsystem—d, r and h—provide these features when implemented with PDMS as the optical material (Supplementary Fig. 7). The combined heights of the microlenses, supporting posts and base membrane (that is, $r + h + t$) position the photodiodes at distances of one focal length ($f = rn/(n - 1) = r + h + t$) from the lens surface. Collimated light at normal incidence focuses to spot sizes that are smaller (about 100 μm × 100 μm) than the areas of the photodiodes.

Retaining these optical parameters throughout the process of transformation from planar to hemispherical shapes represents a challenge that can be addressed with two new design/integration approaches. The first approach involves a method for bonding the optical and electrical subsystems at the positions of the photodiodes/blocking diodes only. This configuration ensures optical alignment during subsequent deformation, but allows free motion of the serpentine interconnects to minimize their effects on the overall mechanics. The resulting response of the system to applied force is dominated by the elastic behaviour of the PDMS (modulus about 1 MPa), and is nearly independent of the hard materials found in the array of photodiodes/blocking diodes (Si, with modulus about 150 GPa, Au, with modulus about 80 GPa and polyimide, with modulus about 5 GPa)[27,28]. In particular, the computed effective modulus of the system is only 1.9 MPa, with global strains that can reach more than 50% in equi-biaxial tension before exceeding the fracture thresholds of the materials.

The second approach exploits a set of dimensional and material choices in the optical subsystem. Here, the modulus of the PDMS is sufficiently small and the heights of the supporting posts are sufficiently large that deformations induced by stretching the base membrane are almost entirely mechanically decoupled from the microlenses. As a result, large strains created by geometry transformation induce no measurable change in the focusing properties. In addition, the combined heights of the microlenses and the posts are large compared to the thickness of the base membrane. This layout minimizes strain at the locations of bonding with the photodiodes/blocking diodes, thereby eliminating the possibility for failure at these interfaces or in the silicon.

Figure 2 summarizes these features in a series of micro X-ray computed tomography (XCT; MicroXCT 400) images and finite element method (FEM) calculations before and after geometrical transformation (see Supplementary Figs 9 and 10 for additional details of FEM and analytical treatments of the mechanics). The results in Fig. 2c highlight four adjacent ommatidia, with strain distributions determined by FEM in each of the different layers of a single ommatidium. The top and bottom surfaces, where the microlenses and photodiodes are located, respectively, show excellent isolation. The peak strains in these regions are <1% (microlenses in box of Fig. 2c) and <0.2% (photodiode/blocking diodes in box of Fig. 2c) even for the large global strains (about 30% or more) that occur in the hemispherical shape. Quantitative analysis of the distribution of r across the entire array, before and after deformation (top panels of Fig. 2d; Supplementary Fig. 12) shows no change, which is consistent with FEM findings (Supplementary Fig. 13). Non-uniform strains lead to a slight, but systematic, spatial variation of $\Delta\Phi$ across the array (bottom left panel of Fig. 2d), as expected based on the mechanics (Supplementary Fig. 13). All ommatidia have an orientation along the direction of the surface normal (that is, the tilt from normal, θ_{tilt}, is zero; bottom right panel of Fig. 2d).

Working apposition cameras formed in this manner have excellent operational characteristics and high yields. Overall image construction follows from a pointwise sampling by the photodiode/blocking diodes of images formed at each microlens. In this way, each ommatidium contributes a single pixel to a different region of the resultant image. Figure 3a schematically illustrates this process through images computed using physically correct ray-tracing procedures (GNU Goptical, see Supplementary Figs 14 and 15 for details) executed in a parallel fashion. Each microlens produces a small image of an object (in this example, a '+' line-art pattern) with a form dictated by the parameters of the lens and the viewing angle (third panel from the left in Fig. 3a). An individual photodiode generates photocurrent only if a portion of the image formed by the associated microlens overlaps the active area. The photodiodes stimulated in this way produce a sampled image (second panel from the left in Fig. 3a) of the object. In biology, rapid motion of the eye and/or the object can yield improvements in effective resolution. Experiments and modelling reported here simulate such effects by scanning the camera from -5.5° to 5.5° in the θ and φ directions with steps of 1.1°. Modelling results appear in the left panel of Fig. 3a (scans from -11° to 11° lead to complete overlap of contributions from neighbouring ommatidia, thereby allowing subtraction of effects of isolated non-functional elements). Figure 3b presents pictures, rendered on hemispherical surfaces with sizes that match the camera, of two different line-art patterns collected using a representative device, for which 166 out of a total of 180 ommatidia function properly (see Supplementary Fig. 16).

Software algorithms and data acquisition systems enable the cameras to adapt to different light levels. When we use a scanning mode for data collection (see Supplementary Information), the results are remarkably consistent with optical modelling that assumes ideal characteristics for the cameras (Fig. 3c). Systematic, quantitative comparisons between parametric simulations and experimentally recorded images indicate correlation values (91.3% and 89.0% in left and right images, respectively) in a range consistent with operation close to limits dictated by the optics and physical designs. Some loss of resolution and edge definition follows from parasitic scattering within the camera. See Supplementary Figs 18 and 19 for details. Other examples appear in Supplementary Fig. 20.

The arthropod eye offers resolution determined by the numbers of ommatidia, and is typically modest (Fig. 3b and c) compared, for example, to mammalian eyes. Two other attributes, however, provide powerful modes of perception. First, the hemispherical apposition design enables exceptionally wide-angle fields of view, without off-axis aberrations. Figure 4a gives an example of this characteristic, through pictures of a line-art soccer ball illustration placed at three different angular positions: -50° (left), 0° (centre), and 50° (right). All three cases show comparable clarity, without anomalous blurring or aberrations, consistent with the proper, independent functioning of ommatidia across the array. Equivalent imaging modes are difficult or impossible to obtain using planar detector technologies even with sophisticated fish-eye lenses, spherical mirrors or other specialized optics. Quantitative analysis can be performed through laser illumination at angles ranging from -80° to 80° with 20° steps along both the x and y directions, as shown in a single composite image in Fig. 4b (see individual images in Supplementary Fig. 21). The uniformity in sizes, shapes, illumination levels and positions of these spots are consistent with expected behaviour over the entire approximately 160° field of view.

The second attribute is the nearly infinite depth of field that results from the short focal length of each microlens and the nature of image formation[3]. In particular, as an object moves away from the camera, the size of the image decreases but remains in focus (Supplementary Fig. 22). A consequence is that the camera can accurately and

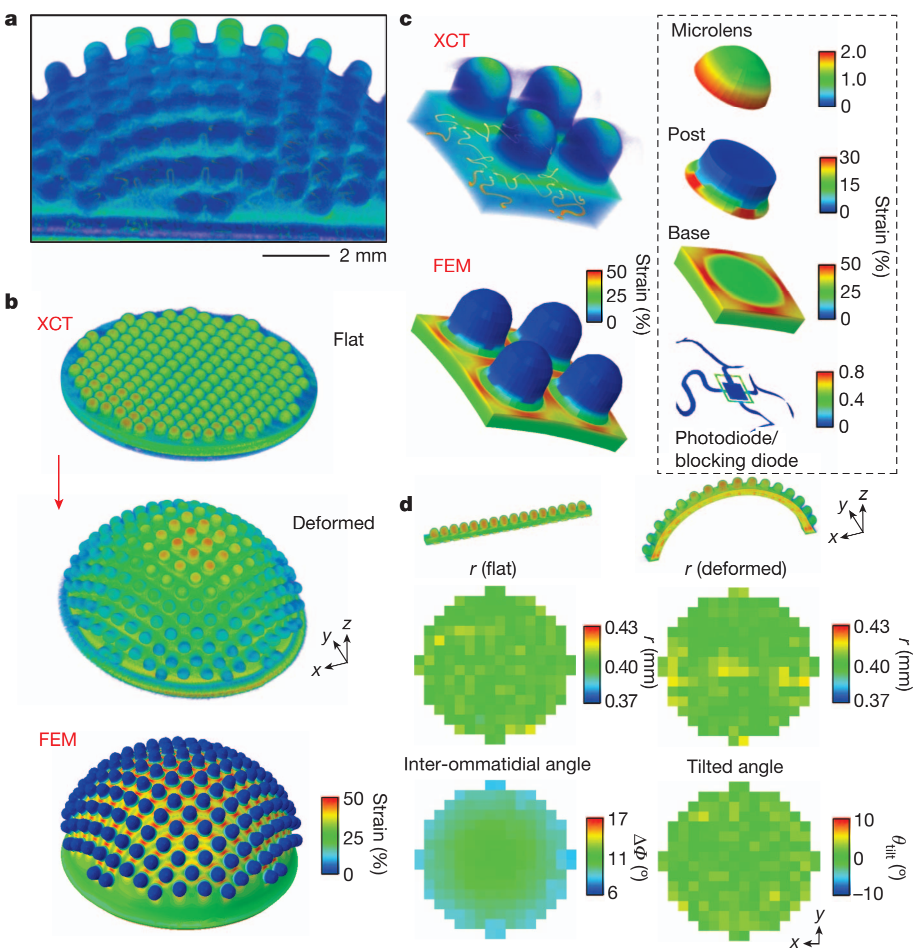


Figure 2 | Computational and experimental studies of the mechanics associated with assembly of a hemispherical, apposition compound eye camera. **a**, XCT image of the imaging component of the camera, showing both the microlenses and the photodiodes/blocking diodes with serpentine interconnects (see Supplementary Fig. 11 for additional details of XCT). **b**, XCT images before (top) and after (middle) elastic deformation into a hemispherical shape, and FEM results for the system after deformation (bottom). **c**, High-resolution XCT image of four adjacent ommatidia located slightly off-axis (in polar and azimuth angles) near the centre of a camera, and FEM computed shape and distributions of strain at this location. The boxed panels highlight strains in different regions of a single ommatidium: microlens, cylindrical post, membrane base and photodiode/blocking diode island with serpentine interconnects. **d**, The top panels are XCT images of 16 microlenses from the middle row of an array in flat (left) and hemispherical (right) geometries. The middle panels are colour maps of the radii of curvature (r) of microlenses in the array, in flat (left) and hemispherical deformed (right) geometries. The bottom panels are colour maps of $\Delta\Phi$ (left) and the angle of tilt of ommatidia away from the surface normal θ_{tilt} (right); both in the hemispherical deformed configuration.

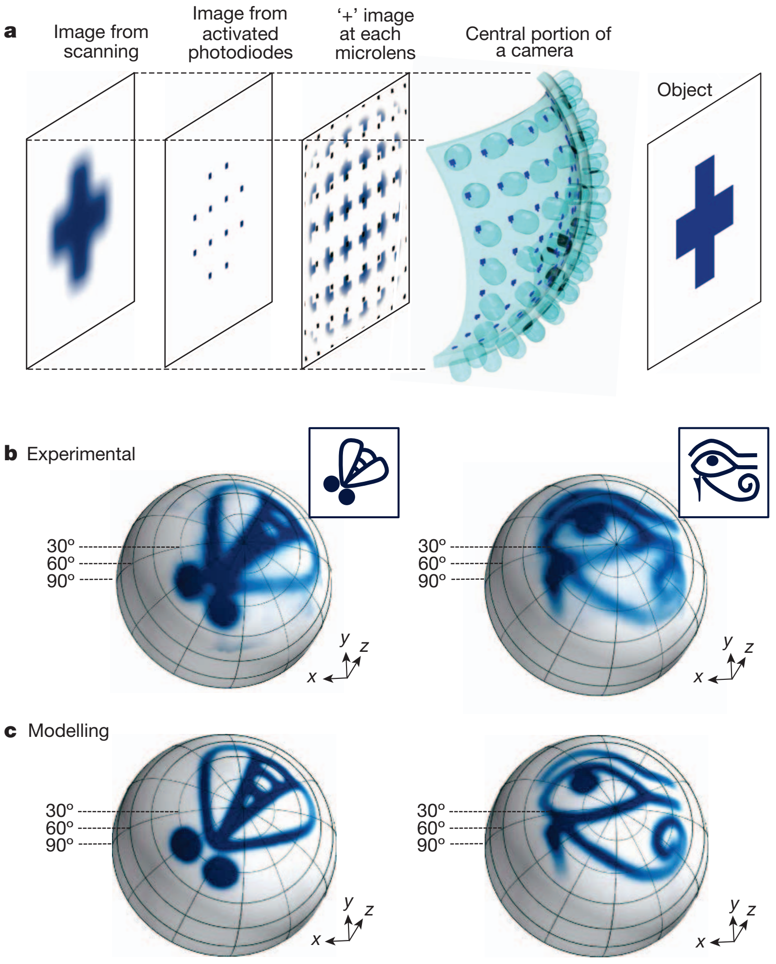


Figure 3 | Operating principles of a hemispherical, apposition compound eye camera and representative pictures. **a**, Conceptual view of image formation, illustrated by quantitative ray-tracing results for the simple case of an 8 × 8 hemispherical array of ommatidia, corresponding to the central region of the camera. Each microlens generates an image of the object ('+' pattern in this example), with characteristics determined by the viewing angle. Overlap of a portion of each image with the active area of a corresponding photodiode (black squares in the third frame from left) generates a proportional photocurrent at this location of the array (second frame from left). The result is a sampled reproduction of the object. Improved effective resolution can be realized by scanning the camera (from −5.5° to 5.5° in the θ and φ directions with 1.1° steps, as shown in the left frame). **b**, Pictures (main panels) of line-art illustrations of a fly (left inset) and a 'Horus eye' (an Egyptian hieroglyphic character) (right inset) captured with a hemispherical, apposition compound eye camera, each rendered on a hemispherical surface that matches the shape of the device. Experimental setups appear in Supplementary Fig. 17. **c**, Images as in **b**, computed by ray-tracing analysis, assuming ideal construction and operation of the camera.

simultaneously render pictures of multiple objects in a field of view, even at widely different angular positions and distances. Figure 4d presents the results of demonstration experiments. Even though movement of the object away from the camera changes its size in the corresponding image, the focus is maintained. Objects with the same angular size that are located at different distances produce images with the same size, all of which is consistent with modelling (Supplementary Fig. 23).

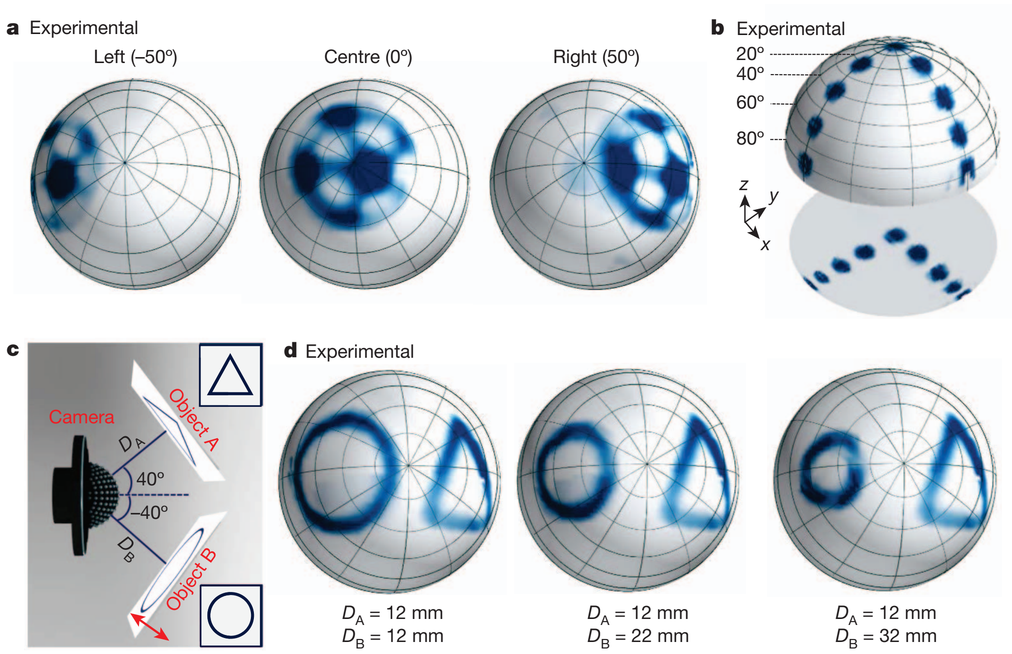


Figure 4 | Imaging characteristics of a hemispherical, apposition compound eye camera. **a**, Pictures of a soccer ball illustration captured at three different polar angles relative to the centre of the camera: −50° (left), 0° (centre), and 50° (right). **b**, Composite picture corresponding to sequential illumination of the camera with a collimated laser beam at a nine different angles of incidence (from −80° to 80° in both x and y directions, with 20° steps), displayed on a hemispherical surface (top) and projected onto a plane (bottom). **c**, Schematic illustration of an experimental setup to demonstrate key imaging characteristics. One object (object A, triangle) lies at an angular position of 40° and distance D_A; the other (object B, circle) at −40° and D_B. **d**, Pictures of these objects collected at different values of D_B.

The cameras reported here incorporate approximately twice as many ommatidia (about 180) as eyes found in some worker ants (about 100 in *Linepithema humile*)[29], significantly fewer than in dragonflies (about 28,000 in *Anax junius*)[10] or praying mantises (about 15,000 in *Stagmatoptera biocellata*)[30], but all with similar fields of view (an estimated 140–180 degrees).

A key defining attribute of our elastomeric optical and deformable electronic subsystems is their applicability to devices with large numbers of ommatidia, diverse spatial layouts, and dimensions into the micrometre regime. Compatibility with silicon technology suggests that commercially available sophistication in imaging arrays and straightforward advances in assembly hardware can enable apposition cameras with resolution and other capabilities that significantly exceed those in known species of arthropods. Specific application requirements and design considerations will dictate the choice between apposition cameras and more conventional approaches that use advanced imaging systems based on fish-eye lenses and others. Other important directions for future research include efforts to expand capabilities beyond those found in biology, such as engineering systems for continuous tuning of the curvature of the hemispherical supporting substrate. Biologically inspired schemes for adapting to different light levels are also of interest.

METHODS SUMMARY

The optical subsystem was formed by casting and curing a prepolymer to PDMS (Sylgard 184, Dow Corning) against a precision, micro-machined aluminium mould and associated fixture. Release of the cured PDMS yielded microlenses and supporting posts with a thin PDMS membrane as a base, each with well defined dimensions. Fabrication of the electrical subsystem involved a series of thin-film processing steps conducted on a silicon-on-insulator wafer. As a final step, etching with concentrated hydrofluoric acid removed the buried oxide layer. Subsequent transfer printing onto the rear plane of the optical subsystem used a homebuilt apparatus, with integrated microscope stage to enable precise alignment. Irreversible bonding at the positions of the photodiodes/blocking diodes was enabled by layers of SiO_2 deposited only in these regions, to allow condensation reactions between their hydroxyl-terminated surfaces and those of the PDMS. A polyimide film with metal contact pads was mounted onto the periphery of the resulting system with an adhesive. A custom mounting assembly and sealed chamber enabled hydraulic deformation from a planar to hemispherical geometry. A hemispherical supporting rod made of PDMS mixed with black silicone pigment (Smooth-on Inc.), and coated with a thin layer of adhesive, held the system in its deformed, hemispherical geometry. A perforated sheet of black silicone, formed by laser machining of thin film membrane (Ecoflex, Smooth-on) mixed with carbon black powder (Strem Chemical), was manually stretched and assembled onto the imager. Mounting on a printed circuit board using mechanical pressure applied with a plastic frame established good electrical contact between the printed circuit board and the metal contacts on the polyimide film. The completed cameras collected pictures of opaque line-art patterns on transparency foils illuminated from behind with diffuse white light, using automated scanning and data acquisition systems.

Received 21 January; accepted 18 March 2013.

1. Warrant, E. & Nillson, D.-E. *Invertebrate Vision* Ch. 1 (Cambridge Univ. Press, 2006).
2. Dudley, R. *The Biomechanics of Insect Flight: Form, Function, Evolution* Ch. 5 (Princeton Univ. Press, 2000).
3. Floreano, D., Zufferey, J.-C., Srinivasan, M. V. & Ellington, C. *Flying Insects and Robot* Ch. 10 (Springer, 2009).
4. Wheeler, W. M. *Ants: their Structure, Development and Behavior* Ch. 4 (Columbia Univ. Press, 1910).
5. Chapman, J. A. Ommatidia numbers and eyes in Scolytid beetles. *Ann. Entomol. Soc. Am.* **65,** 550–553 (1972).
6. Horridge, G. A., McLean, M., Stange, G. & Lillywhite, P. G. A diurnal moth superposition eye with high resolution *Phalaenoides tritifica* (Agaristidae). *Proc. R. Soc. Lond. B* **196,** 233–250 (1977).
7. Kral, K. & Stelzl, M. Daily visual sensitivity pattern in the green lacewing *Chrysoperla carnea* (Neuroptera: Chrysopidae). *Eur. J. Entomol.* **95,** 327–333 (1998).
8. Nilsson, D.-E. A new type of imaging optics in compound eyes. *Nature* **332,** 76–78 (1988).
9. Zeil, J. A new kind of neural superposition eye: the compound eye of male Bibionidae. *Nature* **278,** 249–250 (1979).
10. Land, M. F. & Nillson, D.-E. *Animal Eyes* (Oxford Univ. Press, 2002).
11. Land, M. F. The optics of animal eyes. *Contemp. Phys.* **29,** 435–455 (1988).
12. Nillson, D.-E. Vision optics and evolution. *Bioscience* **39,** 298–307 (1989).
13. Ko, H. C. *et al.* A hemispherical electronic eye camera based on compressible silicon optoelectronics. *Nature* **454,** 748–753 (2008).
14. Jung, I. *et al.* Dynamically tunable hemispherical electronic eye camera system with adjustable zoom capability. *Proc. Natl Acad. Sci. USA* **108,** 1788–1793 (2011).
15. Hung, P. J., Jeong, K., Liu, G. L. & Lee, L. P. Microfabricated suspensions for electrical connections on the tunable elastomer membrane. *Appl. Phys. Lett.* **85,** 6051–6053 (2004).
16. Jeong, K., Kim, J. & Lee, L. P. Biologically inspired artificial compound eyes. *Science* **312,** 557–561 (2006).
17. Dinyari, R., Rim, S.-B., Huang, K., Catrysse, P. B. & Peumans, P. Curving monolithic silicon for nonplanar focal plane array applications. *Appl. Phys. Lett.* **92,** 191114 (2008).
18. Xu, X., Davanco, M., Qi, X. F. & Forrest, S. R. Direct transfer patterning on three dimensionally deformed surfaces at micrometer resolutions and its application to hemispherical focal plane detector arrays. *Org. Electron.* **9,** 1122–1127 (2008).
19. Street, R. A., Wong, W. S. & Lujan, R. Curved electronic pixel arrays using a cut and bend approach. *J. Appl. Phys.* **105,** 104504 (2009).
20. Tanida, J. *et al.* Thin observation module by bound optics (TOMBO): concept and experimental verification. *Appl. Opt.* **40,** 1806–1813 (2001).
21. Duparré, J., Wippermann, F., Dannberg, P. & Reimann, A. Chirped arrays of refractive ellipsoidal microlenses for aberration correction under oblique incidence. *Opt. Express* **13,** 10539–10551 (2005).
22. Li, L. & Yi, A. Y. Development of a 3D artificial compound eye. *Opt. Express* **18,** 18125–18137 (2010).
23. Franceschini, N., Pichon, J. M. & Blanes, C. From insect vision to robot vision. *Phil. Trans. R. Soc. Lond. B* **337,** 283–294 (1992).
24. Afshari, H. *et al.* The PANOPTIC camera: a plenoptic sensor with real-time omnidirectional capability. *J. Sign. Process Syst.* http://dx.doi.org/10.1007/s11265-012-0668-4 (2012).
25. Someya, T. *Stretchable Electronics* (Wiley, 2013).
26. Land, M. F. Visual acuity in insects. *Annu. Rev. Entomol.* **42,** 147–177 (1997).
27. Wang, S. *et al.* Mechanics of curvilinear electronics. *Soft Matter* **6,** 5757–5763 (2010).
28. Lu, C. *et al.* Mechanics of tunable hemispherical electronic eye camera systems that combine rigid device elements with soft elastomers. *J. Appl. Mech.* (in the press).
29. Wild, A. L. *Taxonomic Revision of the Ant Genus* Linepithema *(Hymenoptera: Formicidae)* (Univ. California Press, 2007).
30. Barrós-Pita, J. C. & Maldonado, H. A fovea in the praying mantis eye II. Some morphological characteristics. *Z. Vgl. Physiol.* **67,** 79–92 (1970).

Supplementary Information is available in the online version of the paper.

Acknowledgements The work on integration schemes and mechanical designs was supported by the Defense Advanced Research Projects Agency (DARPA) Nanoelectromechanical /Microelectromechanical Science & Technology (N/MEMS S&T) Fundamentals programme under grant number N66001-10-1-4008 issued by the Space and Naval Warfare Systems Center Pacific (SPAWAR). The work on materials, optical modelling and imaging aspects was supported by the National Science Foundation through an Emerging Frontiers in Research and Innovation (EFRI) programme.

Author Contributions Y.M.S., Y.X., V.M., J.X. and J.A.R. designed the experiments, Y.M.S., Y.X., V.M., J.X., I.J., K.-J.C., Z.L., H.P., C.L., R.-H.K., R.L., K.B.C., Y.H. and J.A.R. performed the experiments and analysis. Y.M.S., V.M., J.X. and J.A.R. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.A.R. (jrogers@illinois.edu).

doi:10.1038/nature12021

Long-term sedimentary recycling of rare sulphur isotope anomalies

Christopher T. Reinhard[1,2], Noah J. Planavsky[1,2] & Timothy W. Lyons[2]

The accumulation of substantial quantities of O_2 in the atmosphere has come to control the chemistry and ecological structure of Earth's surface. Non-mass-dependent (NMD) sulphur isotope anomalies in the rock record[1] are the central tool used to reconstruct the redox history of the early atmosphere. The generation and initial delivery of these anomalies to marine sediments requires low partial pressures of atmospheric O_2 (p_{O_2}; refs 2, 3), and the disappearance of NMD anomalies from the rock record 2.32 billion years ago[1,4] is thought to have signalled a departure from persistently low atmospheric oxygen levels (less than about 10^{-5} times the present atmospheric level) during approximately the first two billion years of Earth's history. Here we present a model study designed to describe the long-term surface recycling of crustal NMD anomalies, and show that the record of this geochemical signal is likely to display a 'crustal memory effect' following increases in atmospheric p_{O_2} above this threshold. Once NMD anomalies have been buried in the upper crust they are extremely resistant to removal, and can be erased only through successive cycles of weathering, dilution and burial on an oxygenated Earth surface. This recycling results in the residual incorporation of NMD anomalies into the sedimentary record long after synchronous atmospheric generation of the isotopic signal has ceased, with dynamic and measurable signals probably surviving for as long as 10–100 million years subsequent to an increase in atmospheric p_{O_2} to more than 10^{-5} times the present atmospheric level. Our results can reconcile geochemical evidence for oxygen production and transient accumulation with the maintenance of NMD anomalies on the early Earth[5–8], and suggest that future work should investigate the notion that temporally continuous generation of new NMD sulphur isotope anomalies in the atmosphere was likely to have ceased long before their ultimate disappearance from the rock record.

One of the most important recent advances in studies of Earth's early atmospheric chemistry has been the demonstration that NMD sulphur isotope anomalies, often of very large magnitude, are preserved in sedimentary sulphide and sulphate minerals more than ~2.32 Gyr old[1,4]. The generation and preservation of these anomalies is generally considered to require active and widespread tropospheric photochemistry involving SO_2 dissociation at short wavelengths, which in turn implies minimal ozone column depth[2]; a strongly reducing atmosphere, such that multiple exit channels for sulphur at different redox states can be maintained[3,9]; and minimal metabolic overprinting of atmospherically derived isotope anomalies within marine environments[10]. The second and third conditions result from simple mass balance: even if NMD anomalies are generated in the atmosphere, isotopically complementary sulphur pools must be removed from the atmosphere and transported to marine sediments with minimal homogenization by inorganic or biological processes. Under these conditions, photochemically derived sulphur containing NMD isotope anomalies will be delivered to the hydrosphere and ultimately buried as a constituent of various sulphur-bearing mineral phases, primarily pyrite (FeS_2). The presence of these anomalies to varying degrees between ~3.8 Gyr ago, the time of the earliest sedimentary record, and ~2.32 Gyr ago is interpreted to reflect a strongly reducing atmosphere over this entire interval, with the implication that atmospheric p_{O_2} was extremely low for more than half of Earth's history (Fig. 1a). Implicit in this framework is the notion that the generation and transfer of these anomalies into the upper crust through the burial of authigenic marine minerals provides an effectively instantaneous record of ambient atmospheric chemistry, but this assumption ignores the potential importance of sedimentary recycling.

There is a striking asymmetry in the $\Delta^{33}S$ record through Archaean time (Fig. 1a, b), with the data skewed in favour of positive $\Delta^{33}S$ values. Importantly, it is the preservation (and associated crustal recycling) of this NMD sulphur isotope asymmetry that allows for the possibility of a temporal lag between the generation and the ultimate removal of the signal from the oceanic sulphur reservoir. We emphasize that although there are probably several mechanistic explanations for this pattern[11–13] (Supplementary Information), what matters foremost for our purposes is the veracity of this empirical observation, regardless of mechanism. This observed asymmetry could be misleading if a sedimentary sulphate reservoir with a complementary negative isotopic composition were deposited synchronous with the generation of the record shown in Fig. 1 but has not been preserved through geologic time as a consequence of more rapid weathering, or if seawater sulphate with the negative $\Delta^{33}S$ complement was thermochemically or microbially reduced and buried into a weatherable sedimentary sulphide reservoir that has been strongly undersampled. In the first case, analysis of the more abundant sulphur isotopes (^{34}S and ^{32}S) indicates that essentially all sulphur entering the Earth surface system was removed as a constituent of pyrite during the Archaean eon[14], leaving little scope for an isotopically complementary Archaean sulphate reservoir that has left no trace on the modern Earth. Consistent with this, extremely low seawater sulphate concentrations during the Archaean relative to the present[15] would probably have rendered large-scale evaporite formation and burial extremely difficult. In the second case, the cumulative average for published $\Delta^{33}S$ values has continued to point to a predominance of positive values in bulk rock measurements as the database of Archaean rare sulphur isotopes has increased in size, while the confidence interval around the mean has continually decreased (Fig. 1b). This relationship indicates that the asymmetry towards positive values does not reflect a sampling bias.

To explore the implications of this asymmetry for long-term recycling of NMD sulphur isotope anomalies, we used a well-established numerical modelling approach. The specific goal was to quantify the importance of recycling $\Delta^{33}S$ signals between the ocean and upper crust (Fig. 2). The model begins with a variation on a class of simple box models used to describe the surface cycling of carbon and sulphur during the Phanerozoic eon, termed 'rapid recycling' models[16–18]. This group of models and our specific approach are variants on models for global carbon–sulphur–oxygen cycling that have been used for decades to explore the dynamics of these coupled biogeochemical cycles at Earth's surface[19,20]. Such models have been used extensively to predict

[1]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91126, USA. [2]Department of Earth Sciences, University of California – Riverside, Riverside, California 92521, USA.

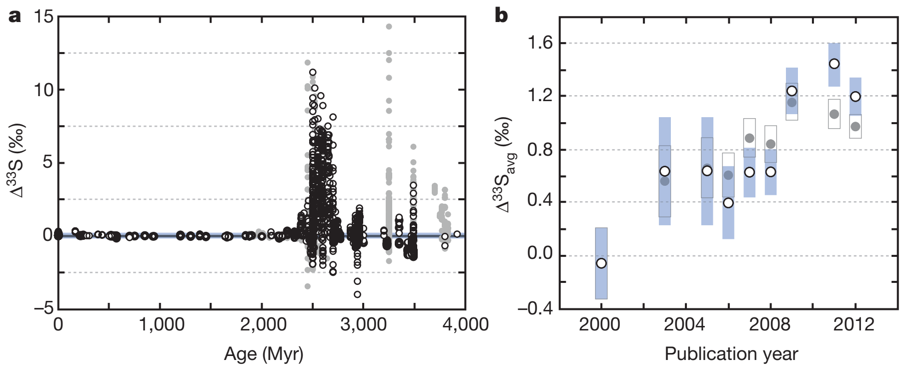


Figure 1 | The rare sulphur isotope record through time. **a**, Data for sedimentary sulphate and sulphide minerals, cast as $\Delta^{33}S$ (where $\Delta^{33}S = \delta^{33}S - 0.515\delta^{34}S$) versus time. The shaded box (0.06±0.16‰) denotes the average ± 2 s.d. of all data from within the past 2,200 Myr. Grey points are data generated through secondary ion mass spectrometry (SIMS), and open circles are bulk rock values. **b**, The cumulative average $\Delta^{33}S$ anomaly as a function of database age for Archaean and early Palaeoproterozoic samples. Grey points show the cumulative average $\Delta^{33}S$ value for the entire database through time, with the open boxes denoting the 95% confidence interval. Open circles show the cumulative average $\Delta^{33}S$ value for the database after removing data generated through SIMS and data from macroscopic and clearly secondary sulphide textures, with the blue boxes denoting the 95% confidence interval. See Supplementary Information for database details.

atmospheric O_2 and CO_2 concentrations that compare well with independent proxy reconstructions of atmospheric composition during the Phanerozoic[21,22]. Our model, which tracks only sulphur, partitions sulphur into three reservoirs: the oceanic sulphate pool and two crustal reservoirs of sedimentary pyrite (Fig. 2). The two crustal reservoirs are referred to as 'young' and 'old', and the primary difference between them other than their overall mass is the speed at which they are recycled. The models build from the geologically reasonable premise that the most recently deposited sediments are more likely to be recycled on a short timescale. Fluxes between reservoirs are predominantly first order with respect to mass; their magnitude depends on the size of the reservoir from which the flux is derived. Three notable exceptions are volcanic inputs and the weathering of igneous (and, thus, isotopically normal, with $\Delta^{33}S \approx 0$‰) sulphides, which are both imposed as constant fluxes within a given model run, and the flux between the two crustal pyrite reservoirs. The latter is set equal to the weathering flux from the old pyrite reservoir such that the mass of this reservoir does not change[16].

Our main interest here is tracking the $\Delta^{33}S$ of seawater sulphate, because this signal will be directly incorporated into sedimentary sulphide minerals under the logical assumption that all subsequent isotope fractionations are mass dependent. We note, however, that there was likely to have been spatial isotopic heterogeneity within the ocean if marine sulphate concentrations were very low, and this possibility has not been incorporated into our model. However, the primary result of spatial isotopic heterogeneity would be to introduce scatter around the trends presented here. In effect, the NMD sulphur isotope signal behaves as a conservative tracer when cycled through a purely mass-dependent sulphur cycle at Earth's surface. In our model, the isotopic composition of seawater sulphate, δ^{3x}_{sw}, will evolve through time according to (Methods and Supplementary Information)

$$M_{sw}\frac{d\delta^{3x}_{sw}}{dt} = \sum_i \left[F_i(\delta^{3x}_i - \delta^{3x}_{sw})\right] - F_{b,py}\Delta^{3x}_{py} \qquad (1)$$

where M_{sw} is the oceanic mass of seawater sulphate, F_i and δ^{3x}_i are respectively the input flux to the ocean and the isotopic composition of reservoir i (from weathering and volcanic sources), $F_{b,py}$ denotes the pyrite burial flux, Δ^{3x}_{py} denotes the isotopic fractionation between seawater sulphate and sedimentary pyrite, and $x = 3$, 4 or 6.

The model tracks all four stable sulphur isotopes and includes a parameterization of biologically induced isotope fractionation, but we restrict our attention here to the $\Delta^{33}S$ composition of sedimentary pyrite formed from chemical or microbial reduction of seawater sulphate, which is derived primarily from the weathering of pyrite. To illustrate most clearly the importance of the sedimentary recycling of $\Delta^{33}S$ signals, we assume at the beginning of each model run that atmospheric p_{O_2} increases instantaneously above values that allow for the generation and preservation of new NMD sulphur anomalies. All isotope fractionations imposed thereafter are mass dependent and are controlled by metabolic fractionation during microbial sulphate reduction, which is parameterized as a function of ambient seawater sulphate concentration (Supplementary Information). For the purposes of illustration, our simulations are initialized with a $\Delta^{33}S$ value for seawater sulphate of −1.0‰ (consistent in sign with expectations from photochemical experiments[2] and analyses of Archaean sulphates[1,23,24] and seafloor sulphide minerals[9,23]), although we emphasize that for many periods of Archaean time this value was considerably more negative (Fig. 1a). We then study the effect of various values for the initial $\Delta^{33}S$ of rapidly weathering sedimentary pyrite, isotopically normal input fluxes (volcanic and igneous weathering) and seawater sulphate concentrations. We stress, however, that these values and the parameter space explored, although certainly plausible, are only meant to be delineative. Our goal is to investigate the timescales on which crustal recycling of NMD isotope anomalies can be expected to leave an imprint on the isotopic composition of the oceanic sulphate reservoir, rather than to simulate specific features of the sedimentary NMD sulphur record. The latter goal is a subject for future research.

There is a notable 'memory effect' associated with the sedimentary recycling of $\Delta^{33}S$ anomalies (Fig. 3a–d), and this memory effect

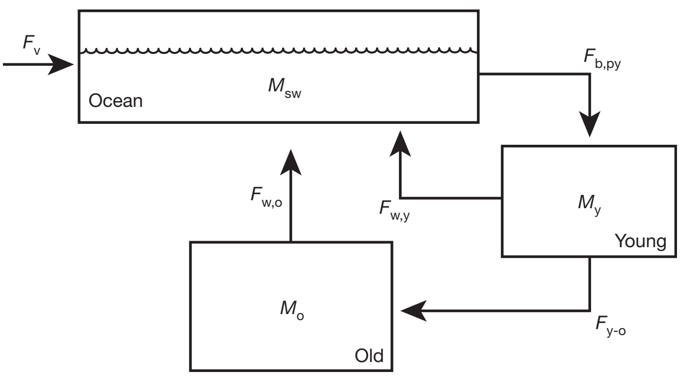


Figure 2 | Schematic diagram of the sulphur isotope mass balance model. Arrows denote flux terms (F_i, where in addition to fluxes already described, $F_{y\text{-}o}$ represents the aging flux of the young crustal reservoir), whereas boxes denote various oceanic and crustal sulphur reservoirs (M_{sw}, M_y and M_o are the sizes of the seawater reservoir and the two crustal reservoirs, respectively) (Methods Summary).

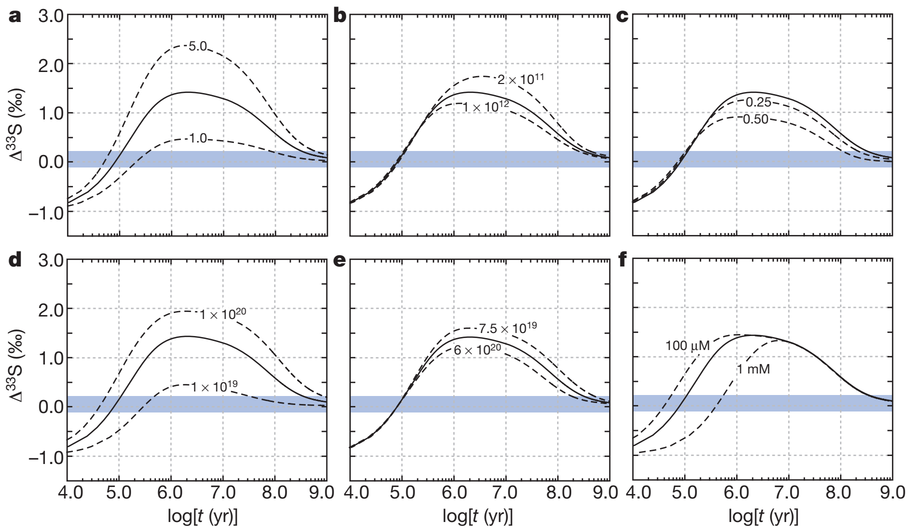


Figure 3 | Modelled changes to the Δ^{33}S value of seawater sulphate after the onset of oxidative sulphur cycling. The black solid curve in all cases represents the reference model (Supplementary Information). Dashed curves represent sensitivity analyses as follows: increasing and decreasing the Δ^{33}S of the rapidly weathering pyrite reservoir by 2‰ (**a**); a range of volcanic sulphur fluxes between 2×10^{11} mol yr^{-1} and 1×10^{12} mol yr^{-1} (**b**); increasing the fractional contribution of igneous sulphide weathering to total weathering (**c**) (values approaching or exceeding 0.5 are considered extremely unlikely; see Supplementary Information); increasing and decreasing, by factors of 2 and 5, respectively, the initial size of the rapidly weathering pyrite reservoir (**d**); increasing and decreasing, by factors of 2 and 5, respectively, the size of the slowly weathering pyrite reservoir (**e**); varying initial oceanic sulphate concentration between 100 μM and 1 mM (**f**). The shaded box denotes the average ± 2 s.d. of all data from within the past 2,200 Myr. Note that time is displayed on a logarithmic scale.

is difficult to avoid in the parameter space that we consider reasonable. For example, a sizable decrease in the magnitude of residual Δ^{33}S values can be achieved by decreasing by a factor of five the size of the rapidly weathering pyrite reservoir from that in the reference model (Fig. 3d). However, this results in geologically unrealistic fluxes and residence times relative to observed mass–age and area–age distributions of weatherable sedimentary rocks[25,26] and the timescales of cycling through the Earth surface sulphur reservoir[27]. In other words, this decrease would require a severe departure from the sedimentary rock cycle that has been in place for most of Earth's history (Supplementary Information). Similarly, large, mass-dependent sulphur fluxes from volcanic or igneous sources can in principle dilute the memory effect, but extreme values for these fluxes are required for pronounced attenuation of the residual isotopic signal (Fig. 3b, c). Once NMD isotope fractionations have been introduced into the system in an asymmetric fashion, the best taphonomic conditions for their preservation, in fact, result from all isotope fractionations being mass dependent. This mass dependence prevents homogenization, either locally or on a broad spatial scale, through mixing with isotopically complementary pools.

A residual NMD isotope signal incorporated into sedimentary rocks can persist in our model for roughly 10^8 yr after the cessation of its atmospheric production. However, it is difficult to extend the memory effect beyond the order of a single Wilson cycle (that is, ~200–250 Myr; Fig. 3). We note, however, that our simple model does not account for time-dependent changes in volcanic sulphur fluxes and microbial processing, and that large perturbations to either of these parameters during the pulsed decay of a residual NMD signal could in principle shorten this timescale somewhat. In addition, the texture, or temporal pattern, of the signal decay (regardless of the ultimate timescale of the memory effect) will depend strongly on the initial Δ^{33}S value and the initial size of the seawater sulphate reservoir (Fig. 3a, f). Larger seawater sulphate concentrations will result in greater temporal inertia as the system moves towards the isotopic properties of the weathering input with positive Δ^{33}S values.

An important outcome of this dynamic within the residual isotopic signal is that oscillations in atmospheric p_{O_2} near or well above the threshold for the generation and synchronous preservation of NMD isotope anomalies may be expected to produce a wide range of temporal responses depending on the speed and periodicity of the oscillation. Most of the texture during each simulation is centred on the 10^5–10^7-yr timescale, during which the Δ^{33}S values of seawater sulphate in the model will change from being strongly negative to being strongly positive (Fig. 3). In addition, during periods of NMD sulphur production the Δ^{33}S composition of sea water should rapidly change back to values governed by atmospheric sulphur input, on a timescale ultimately dependent on the size of the seawater sulphate reservoir but probably not greatly in excess of ~10^5 yr. The net result of this would be an extremely dynamic, perhaps even noisy, rare sulphur isotope record, despite extremely long periods of oxidative crustal weathering.

It is commonly argued that biological oxygen production preceded the broad-scale and effectively permanent accumulation of oxygen in Earth's atmosphere by perhaps 200 Myr or more[5–8,28,29]. Consistent with oxygenation before the so-called Great Oxidation Event, some data imply oxidative weathering over this interval[5–8]. Our model suggests the possibility of excursions in atmospheric O_2 content that may have been well above the upper threshold for generating and preserving NMD sulphur isotope anomalies in marine sediments long before the disappearance of these signals from the record, on timescales more than adequate to support extensive oxidative weathering of crustal minerals[5,7]. Alternation between periods of generation and non-generation of atmospheric NMD isotope anomalies on a range of timescales (10^5–10^8 yr) would not be immediately manifest in the removal of these signals from the rock record due to the recycling effect. It has been suggested that the behaviour of O_2 in a relatively reducing atmosphere is likely to be characterized by strong hysteresis[30]—such that the attainment of relatively 'high' O_2 (above ~10^{-5} times the present atmospheric level) may not be readily undone. Nevertheless, with atmospheric p_{O_2} values (and, thus, residence times) far below those characteristic of the modern Earth we would naturally expect high-frequency oscillations in atmospheric p_{O_2} with corresponding periods much shorter than the duration of the crustal memory effect within the Δ^{33}S record, and with biological oxygen production emergent and the interplay between this process and inorganic buffering mechanisms varying in ways that may have been episodic.

The model permits oscillatory behaviour for the oxygen cycle during earlier portions of Archaean time, because NMD signals would persist even as p_{O_2} rose and fell on timescales of millions of years. However,

our model also implies that a pulsed or irreversible rise in atmospheric O_2 during the late Archaean may have preceded the ultimate disappearance of NMD sulphur isotope anomalies from the rock record by $\sim 10^7$–10^8 yr. Combined, these results suggest that the texture of atmospheric redox evolution on the early Earth may have been highly dynamic and may call into question the notion of a Great Oxidation Event as strictly understood—that is, that there was a moment or brief period in Earth's history when the oxygen concentration increased permanently to levels above those required to support the production and preservation of NMD anomalies in the atmosphere—and instead suggest that the oxygenation of the atmosphere could have been a protracted process[8]. Viewed in this way, our model suggests that many of the climatological and geochemical upheavals witnessed by the Archaean–Proterozoic transition, including the earliest recorded widespread glaciations[31] and the deposition of perhaps the largest iron and manganese deposits in Earth's history[32], may have been linked more directly to excursions in atmospheric O_2 content than current interpretations of the rare sulphur isotope record indicate. The details of oscillatory redox behaviour and the timing of oxygen's irreversible increase, along with the further constraint of the input parameters controlling the fabric and lags in the NMD record, are topics for future research. Nevertheless, recycling of crustal sulphur with relict NMD isotope anomalies must be considered in further attempts to explore quantitatively the palaeoenvironmental and palaeobiological implications of the Archaean sulphur isotope record.

METHODS SUMMARY

We model the dynamics of the isotopic composition of seawater sulphate through time, as governed by the input fluxes and isotopic compositions of sulphur associated with weathering of sulphides from two sedimentary reservoirs ($F_{w,y}$ and $F_{w,o}$), weathering of sulphide from an igneous (and, thus, isotopically normal) reservoir ($F_{w,ig}$) and volcanic sulphur emissions (F_v), balanced against the removal of sulphur from the ocean in association with the burial of sedimentary pyrite ($F_{b,py}$). The time-dependent isotope mass balance equation for seawater sulphate is then given by equation (1). The isotopic composition of a given reservoir i is defined according to conventional 'delta' notation as $\delta_i^{3x} = \left((^{3x}S/^{32}S)_{sample}/(^{3x}S/^{32}S)_{standard} - 1 \right) \times 10^3$‰, where the standard is defined relative to the isotopic composition of Vienna Canyon Diablo Troilite. A detailed discussion of the model set-up and parameterization, references for the sulphur isotope database and details of compilation and statistical treatment are provided in Methods and Supplementary Information.

Full Methods and any associated references are available in the online version of the paper.

Received 8 August 2012; accepted 19 February 2013.
Published online 24 April 2013.

1. Farquhar, J., Bao, H. & Thiemens, M. Atmospheric influence of Earth's earliest sulfur cycle. *Science* **289,** 756–758 (2000).
2. Farquhar, J., Savarino, J., Airieau, S. & Thiemens, M. H. Observation of wavelength-sensitive mass-independent sulfur isotope effects during SO_2 photolysis: implications for the early atmosphere. *J. Geophys. Res.* **106,** 32829–32839 (2001).
3. Pavlov, A. A. & Kasting, J. F. Mass-independent fractionation of sulfur isotopes in Archean sediments: strong evidence for an anoxic Archean atmosphere. *Astrobiology* **2,** 27–41 (2002).
4. Bekker, A. *et al.* Dating the rise of atmospheric oxygen. *Nature* **427,** 117–120 (2004).
5. Anbar, A. D. *et al.* A whiff of oxygen before the Great Oxidation Event? *Science* **317,** 1903–1906 (2007).
6. Frei, R., Gaucher, C., Poulton, S. W. & Canfield, D. E. Fluctuations in Precambrian atmospheric oxygenation recorded by chromium isotopes. *Nature* **461,** 250–253 (2009).
7. Reinhard, C. T., Raiswell, R., Scott, C., Anbar, A. D. & Lyons, T. W. A late Archean sulfidic sea stimulated by early oxidative weathering of the continents. *Science* **326,** 713–716 (2009).
8. Konhauser, K. O. *et al.* Aerobic bacterial pyrite oxidation and acid rock drainage during the Great Oxidation Event. *Nature* **478,** 369–373 (2011).
9. Ono, S. *et al.* New insights into Archean sulfur cycle from mass-independent sulfur isotope records from the Hamersley Basin, Australia. *Earth Planet. Sci. Lett.* **213,** 15–30 (2003).
10. Halevy, I., Johnston, D. T. & Schrag, D. P. Explaining the structure of the Archean mass-independent sulfur isotope record. *Science* **329,** 204–207 (2010).
11. Farquhar, J. *et al.* Inclusions in diamond and sulfur recycling on early Earth. *Science* **298,** 2369–2372 (2002).
12. Farquhar, J. & Wing, B. A. Multiple sulfur isotopes and the evolution of the atmosphere. *Earth Planet. Sci. Lett.* **213,** 1–13 (2003).
13. Bekker, A. *et al.* Atmospheric sulfur in Archean komatiite-hosted nickel deposits. *Science* **326,** 1086–1089 (2009).
14. Canfield, D. E. & Farquhar, J. Animal evolution, bioturbation, and the sulfate concentration of the oceans. *Proc. Natl Acad. Sci. USA* **106,** 8123–8127 (2009).
15. Habicht, K. S., Gade, M., Thamdrup, B., Berg, P. & Canfield, D. E. Calibration of sulfate levels in the Archean ocean. *Science* **298,** 2372–2374 (2002).
16. Berner, R. A. Models for carbon and sulfur cycles and atmospheric oxygen: application to Paleozoic geologic history. *Am. J. Sci.* **287,** 177–196 (1987).
17. Berner, R. A. & Canfield, D. E. A new model for atmospheric oxygen over Phanerozoic time. *Am. J. Sci.* **289,** 333–361 (1989).
18. Berner, R. A. Modeling atmospheric O_2 over Phanerozoic time. *Geochim. Cosmochim. Acta* **65,** 685–694 (2001).
19. Garrels, R. M. & Lerman, A. Coupling of the sedimentary sulfur and carbon cycles – an improved model. *Am. J. Sci.* **284,** 989–1007 (1984).
20. Kump, L. R. & Garrels, R. M. Modeling atmospheric O_2 in the global sedimentary redox cycle. *Am. J. Sci.* **286,** 337–360 (1986).
21. Royer, D. L., Berner, R. A. & Beerling, D. J. Phanerozoic atmospheric CO_2 change: evaluating geochemical and paleobiological approaches. *Earth Sci. Rev.* **54,** 349–392 (2001).
22. Scott, A. C. & Glasspool, I. J. The diversification of Paleozoic fire systems and fluctuations in atmospheric oxygen concentration. *Proc. Natl Acad. Sci. USA* **103,** 10861–10865 (2006).
23. Farquhar, J. & Wing, B. A. The terrestrial record of stable sulphur isotopes: a review of the implications for evolution of Earth's sulphur cycle. *Spec. Publ. Geol. Soc. (Lond.)* **248,** 167–177 (2005).
24. Ueno, Y., Ono, S., Rumble, D. & Maruyama, S. Quadruple sulfur isotope analysis of ca. 3.5 Ga Dresser Formation: new evidence for microbial sulfate reduction in the early Archean. *Geochim. Cosmochim. Acta* **72,** 5675–5691 (2008).
25. Bluth, G. J. S. & Kump, L. R. Phanerozoic paleogeology. *Am. J. Sci.* **291,** 284–308 (1991).
26. Gregor, C. B. The mass-age distribution of Phanerozoic sediments. *Geol. Soc. Lond. Mem.* **10,** 284–289 (1985).
27. Garrels, R. M. & Mackenzie, F. T. A quantitative model for the sedimentary rock cycle. *Mar. Chem.* **1,** 27–41 (1972).
28. Kaufman, A. J. *et al.* Late Archean biospheric oxygenation and atmospheric evolution. *Science* **317,** 1900–1903 (2007).
29. Garvin, J., Buick, R., Anbar, A. D., Arnold, G. L. & Kaufman, A. J. Isotopic evidence for an aerobic nitrogen cycling in the latest Archean. *Science* **323,** 1045–1048 (2009).
30. Goldblatt, C., Lenton, T. M. & Watson, A. J. Bistability of atmospheric oxygen and the Great Oxidation. *Nature* **443,** 683–686 (2006).
31. Evans, D. A., Beukes, N. J. & Kirschvink, J. L. Low-latitude glaciation in the Palaeoproterozoic era. *Nature* **386,** 262–266 (1997).
32. Maynard, J. B. The chemistry of manganese ores through time: a signal of increasing diversity of Earth-surface environments. *Econ. Geol.* **105,** 535–552 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements Funding from NSF-EAR and the NASA Exobiology Program supported this research. C.T.R. acknowledges support from an O. K. Earl Postdoctoral Fellowship in Geological and Planetary Sciences at the California Institute of Technology. N.J.P. acknowledges support from NSF-EAR-PDF. Comments and criticism from L. Kump, B. Wing, A. Bekker and K. Konhauser greatly improved the manuscript.

Author Contributions C.T.R. and N.J.P. designed the model. C.T.R. compiled the sulphur isotope database and performed the modelling and statistical analyses. C.T.R. and N.J.P. wrote the manuscript, with contributions from T.W.L.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.T.R. (reinhard@caltech.edu).

METHODS

Model structure. The model consists of three Earth surface sulphur reservoirs: an oceanic sulphate reservoir (M_{sw}) and two crustal sulphur reservoirs (referred to, following refs 16–18, 33, 34, as 'young' (M_y) and 'old' (M_o)). The distinction between two crustal reservoirs of varying cycling speeds was initially introduced to more directly couple the carbon and sulphur isotope compositions of fluxes out of the ocean to that of fluxes into the ocean, in an effort to alleviate physically unrealistic shifts in atmospheric composition due to changes in measured isotope ratios of sedimentary carbonate and sulphate minerals[16,33]. However, there is also ample geological justification for such a model configuration[16–18,27,33–35], and subsequent work has shown that this assumption results in a good agreement between proxy-based reconstructions of Phanerozoic atmospheric composition and those derived from mass balance models[21,22,34,36,37].

Model equations. Full derivation of the mass balance equations is given in Supplementary Information. The model solves a series of coupled isotope mass balance equations for oceanic and crustal sulphur:

$$M_{sw}\frac{d\delta_{sw}^{3x}}{dt}=\sum_i\left[F_i(\delta_i^{3x}-\delta_{sw}^{3x})\right]-F_{b,py}\Delta_{py}^{3x}$$

$$M_y\frac{d\delta_y^{3x}}{dt}=F_{b,py}(\delta_{py}^{3x}-\delta_y^{3x})$$

$$M_o\frac{d\delta_o^{3x}}{dt}=F_{y\text{-}o}(\delta_y^{3x}-\delta_o^{3x})$$

where F terms denote sulphur fluxes and M terms refer to reservoir sizes. The δ terms denote sulphur isotope compositions of the different reservoirs and fluxes according to conventional 'delta' notation, $\delta_i^{3x}=\left((^{3x}S/^{32}S)_{sample}/(^{3x}S/^{32}S)_{standard}-1\right)\times 10^3$‰, where the standard is defined relative to the isotopic composition of Vienna Canyon Diablo Troilite, and $x = 3, 4$ or 6. The Δ term in the first equation describes the metabolic fractionation imparted by dissimilatory sulphate reduction as expressed on a global scale, and is described by a Monod-type function (Supplementary Information).

Initial parameterization. Model parameters for our reference case are shown in Supplementary Table 1. Parameter values for the reference case were chosen to approximately satisfy known constraints on the overall size of the crustal sulphur reservoir[16,19,38,39]; the residence time of sulphur as it cycles through the exogenic system[20,27]; the fraction of overall sulphur input derived from the rapidly recycling sulphur reservoir[25,34,40–42]; the residence time of sulphur in the rapidly recycling reservoir with respect to weathering[16,17]; and the residence time of sulphur in the rapidly recycling reservoir with respect to removal to the old reservoir[16,34,42] (the 'aging flux' of the young pyrite reservoir). The range of F_v values was chosen to encompass estimates of the modern volcanic sulphur flux and values scaled up to reflect the possibility of greater crustal heat flow and volcanic activity during Earth's early history. Estimates of the modern volcanic sulphur flux are typically of the order of $\sim(2–3)\times10^{11}$ mol yr^{-1} (refs 39, 43–47), and we use an estimate of 2×10^{11} mol yr^{-1} as our low volcanic flux. Heat flow through the crust has decreased with time, and as a result it is typically assumed that Earth's early history was characterized by increased rates of volcanism. Estimates vary, but it is unlikely that crustal heat flow during the Archaean was more than ~3–4 times that of the modern Earth[48–51]. We therefore use a volcanic sulphur input of 1×10^{12} mol yr^{-1} as our high volcanic flux. We note, however, that higher rates of heat flow through the crust need not require an increase in the mass flux from subaerial volcanic activity—much of this heat loss may have been accommodated by submarine mafic spreading centres[52], which are essentially sulphur neutral in an anoxic and iron-buffered deep ocean. In each model run, the pyrite burial rate constant (k_{py}) is solved for to attain a steady state with the prescribed volcanic sulphur flux. Oxidative weathering of sedimentary and igneous rocks is then initialized, and the model allowed to evolve freely.

33. Berner, R. A. Biogeochemical cycles of carbon and sulfur and their effect on atmospheric oxygen over Phanerozoic time. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **75,** 97–122 (1989).
34. Berner, R. A. GEOCARBSULF: a combined model for Phanerozoic atmospheric O_2 and CO_2. *Geochim. Cosmochim. Acta* **70,** 5653–5664 (2006).
35. Garrels, R. M. & Mackenzie, F. T. Sedimentary rock types: relative proportions as a function of geological time. *Science* **163,** 570–571 (1969).
36. Royer, D. L., Berner, R. A., &. Park, J. Climate sensitivity constrained by CO_2 concentrations over the past 420 million years. *Nature* **446,** 530–532 (2007).
37. Berner, R. A. & Kothavala, Z. GEOCARB III: a revised model of atmospheric CO_2 over Phanerozoic time. *Am. J. Sci.* **301,** 182–204 (2001).
38. Li, Y.-H. Geochemical mass balance among lithosphere, hydrosphere, and atmosphere. *Am. J. Sci.* **272,** 119–137 (1972).
39. Canfield, D. E. The evolution of the Earth surface sulfur reservoir. *Am. J. Sci.* **304,** 839–861 (2004).
40. Garrels, R. M., Lerman, A. & Mackenzie, F. T. Controls of atmospheric O_2 and CO_2: past, present, and future. *Am. Sci.* **64,** 306–315 (1976).
41. Blatt, H. & Jones, R. L. Proportions of exposed igneous, metamorphic, and sedimentary rocks. *Geol. Soc. Am. Bull.* **86,** 1085–1088 (1975).
42. Berner, R. A. A model for atmospheric CO_2 over Phanerozoic time. *Am. J. Sci.* **291,** 339–376 (1991).
43. Berresheim, H. & Jaeschke, W. The contribution of volcanoes to the global atmospheric sulfur budget. *J. Geophys. Res.* **88,** 3732–3740 (1983).
44. Walker, J. C. G. & Brimblecombe, P. Iron and sulfur in the pre-biologic ocean. *Precambr. Res.* **28,** 205–222 (1985).
45. Andres, R. J. & Kasgnoc, A. D. A time-averaged inventory of subaerial volcanic sulfur emissions. *J. Geophys. Res.* **103,** 25251–25261 (1998).
46. Halmer, M. M., Schmincke, H.-U. & Graf, H.-F. The annual volcanic gas input into the atmosphere, in particular to the stratosphere: a global data set for the past 100 years. *J. Volcanol. Geotherm. Res.* **115,** 511–528 (2002).
47. Canfield, D. E., Rosing, M. T. & Bjerrum, C. Early anaerobic metabolisms. *Phil. Trans. R. Soc. B* **361,** 1819–1836 (2006).
48. Schubert, G., Stevenson, D. & Cassen, P. Whole planet cooling and the radiogenic heat source contents of the Earth and Moon. *J. Geophys. Res.* **85,** 2531–2538 (1980).
49. Christensen, U. R. Thermal evolution models for the Earth. *J. Geophys. Res.* **90,** 2995–3007 (1985).
50. Lenardic, A. Continental growth and the Archean paradox. *Geophys. Monogr. Ser.* **164,** 33–45 (2006).
51. Padhi, C. M., Korenaga, J. & Ozima, M. Thermal evolution of Earth with xenon degassing: a self-consistent approach. *Earth Planet. Sci. Lett.* **341–344,** 1–9 (2012).
52. Kump, L. R. & Barley, M. E. Increased subaerial volcanism and the rise of atmospheric oxygen 2.5 billion years ago. *Nature* **448,** 1033–1036 (2007).

doi:10.1038/nature12059

Linking the evolution of body shape and locomotor biomechanics in bird-line archosaurs

Vivian Allen[1,2], Karl T. Bates[3], Zhiheng Li[4,5] & John R. Hutchinson[2]

Locomotion in living birds (Neornithes) has two remarkable features: feather-assisted flight, and the use of unusually crouched hindlimbs for bipedal support and movement. When and how these defining functional traits evolved remains controversial[1–8]. However, the advent of computer modelling approaches and the discoveries of exceptionally preserved key specimens now make it possible to use quantitative data on whole-body morphology to address the biomechanics underlying this issue. Here we use digital body reconstructions to quantify evolutionary trends in locomotor biomechanics (whole-body proportions and centre-of-mass position) across the clade Archosauria. We use three-dimensional digital reconstruction to estimate body shape from skeletal dimensions for 17 archosaurs along the ancestral bird line, including the exceptionally preserved, feathered taxa *Microraptor*, *Archaeopteryx*, *Pengornis* and *Yixianornis,* which represent key stages in the evolution of the avian body plan. Rather than a discrete transition from more-upright postures in the basal-most birds (Avialae) and their immediate outgroup deinonychosauria[5,6], our results support hypotheses of a gradual, stepwise acquisition of more-crouched limb postures across much of theropod evolution[1–4], although we find evidence of an accelerated change within the clade Maniraptora (birds and their closest relatives, such as deinonychosaurs). In addition, whereas reduction of the tail is widely accepted to be the primary morphological factor correlated with centre-of-mass position and, hence, evolution of hindlimb posture[1–8], we instead find that enlargement of the pectoral limb and several associated trends have a much stronger influence. Intriguingly, our support for the onset of accelerated morpho-functional trends within Maniraptora is closely correlated with the evolution of flight. Because we find that the evolution of enlarged forelimbs is strongly linked, via whole-body centre of mass, to hindlimb function during terrestrial locomotion, we suggest that the evolution of avian flight is linked to anatomical novelties in the pelvic limb as well as the pectoral.

Terrestrial animals exert a force against the ground to support and move their body. The vector of the incurred ground reaction force (GRF) generally points at or close to the centre of mass (CoM) to stabilize the body[9,10]. The GRF is mainly vertical during the middle of the supportive (stance) phase of locomotion (see, for example, refs 11, 12). Bipedal animals such as birds and many extinct non-avian dinosaurs use a single supporting limb for most of the stance phase. Therefore, the foot of this limb must be placed directly underneath the CoM around mid-stance to exert a vertical GRF, and the joints of the limb must be suitably positioned to allow the antigravity muscles to push against the ground (the GRF passes on the flexor side of the ankle, knee and hip[11,13–16]). The location of the CoM is therefore a major determinant of the limb orientation at mid-stance. Hence, the 'crouched' mid-stance postures of Neornithes, in which the hip is highly flexed, placing the feet well cranial to the hip and the knee cranial to the GRF, are correlated with a strongly cranial (for a biped) CoM[8,17]. In contrast, the ancestral archosaur is likely to have had a more caudal CoM[18] and, by inference, a different limb orientation.

Reconstruction of evolutionary trends in CoM position along the bird line therefore represents an important and under-used source of data on the origin and evolution of aspects of pelvic limb function that were inherited by extant birds. Analysis has previously been limited to qualitative inferences of mass distribution from theropod skeletal proportions, which have led to conflicting interpretations of CoM evolution. On the basis of a trend towards reduced tail size along the bird line, it has been suggested that the CoM steadily moved cranially from coelurosaurian theropods to extant birds[1,2]. The inference of a gradual change in pelvic limb posture is supported by contemporaneous trends in hip anatomy indicating increasingly flexed hip joints[3]. Alternatively, it has been suggested that a trend towards a more triangular chest (concentrating chest mass caudally) in theropods closely related to birds counteracted tail reduction to some extent, and that a more concentrated cranial shift in CoM occurred subsequently within the avian stem clade Avialae[6]. Some support for a later, more discrete shift in limb posture and function is intimated by studies finding distinct differences between the pelvic limb proportions[19] and stride parameters[20] of non-avian theropods and extant Avialae. Thus, when and how critical functional traits of living birds evolved remains controversial, and this limited understanding prohibits tests of the interplay between the evolution of terrestrial locomotion and flight, in addition to other physiological and ecological aspects of the origin of birds.

Here we present a quantitative analysis of bird-line CoM evolution, using empirically validated[18] three-dimensional computational models of mass distribution (Methods Summary and Supplementary Video 1) based on digitized fossil specimens of the range of bird-line taxa shown in Fig. 1 (for full specimen data, see Supplementary Table 1; for animated visualizations of all models, see Supplementary Video 2). Representative modelled body volumes are shown in Fig. 2. To address trends along the bird line itself, rather than at terminal taxa, estimates of CoM and other mass properties were mapped onto the evolutionary splitting events, or nodes (Fig. 1; 1–16), using a squared-change parsimony method based on temporal branch length (see Methods Summary). Our results corroborate a significant ($P < 0.05$, $R = 0.44$, Pearson's correlation of phylogenetic node date and CoM estimates) cranial shift in CoM position over the entire bird line. Visualization of the results indicates that this cranial shift was not evenly distributed or monotonic, but started sometime during the diversification of the clade Maniraptora (Fig. 3, between nodes 11 and 12) in the Jurassic period. We also discern a marked cranial shift in CoM position (approximately twofold) that reaches a maximum in basal Ornithurae (birds closely related to Neornithes; Fig. 3, node 15) before shifting somewhat caudally again in Neornithes. Our sensitivity analysis (Fig. 3 error bars; see Methods Summary) indicates that these trends are still evident when allowing for considerable variation in the morphological assumptions underlying our reconstruction methodology.

Figure 4 (black dashed line) shows evolutionary trends in the (size-normalized) first mass moment of individual segments about the mediolateral axis (that is, segment mass multiplied by segment CoM

[1]Institut für Spezielle Zoologie und Evolutionsbiologie, Friedrich-Schiller-University Jena, 07743 Jena, Germany. [2]Structure & Motion Laboratory, The Royal Veterinary College, Hatfield AL9 7TA, UK. [3]Department of Musculoskeletal Biology II, Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool L69 3GA, UK. [4]Key Laboratory of Evolutionary Systematics of Vertebrates, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100034, China. [5]Department of Geological Sciences, University of Texas at Austin, Austin, Texas 78712-1692, USA.

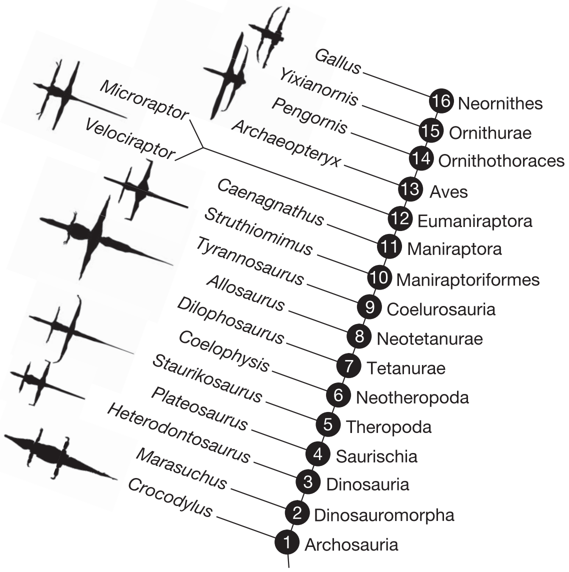


Figure 1 | Dorsal views of models of study taxa, arranged along a phylogenetic tree of Archosauria. Limbs are splayed laterally to show relative dimensions. See Supplementary Tables 1 and 2 for specimen numbers and references used to construct the phylogeny, divergence times and branch lengths.

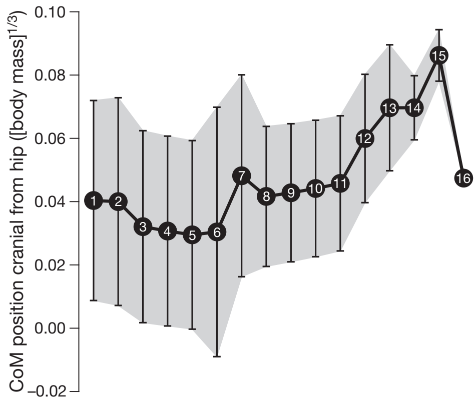


Figure 3 | Estimated evolutionary trends in whole-body CoM position along the craniocaudal axis. Whisker plot bars represent the range of values indicated by our sensitivity analysis. Numbers 1–16 correspond to the nodes in Fig. 1. Values towards the extreme ends of the whisker plots are less plausible, maximal/minimal models with extreme proportions, whereas values towards the middle are more plausible, conservatively proportioned models. Although biomechanically implausible, whisker plot ranges showing negative craniocaudal CoM positions (behind the hip) are included for completeness.

position along the craniocaudal axis), representing the total influence of each segment on whole-body craniocaudal CoM position (see equation (1) in Methods Summary). Positive shifts concurrent, and therefore potentially correlated, with the Maniraptora-to-Ornithurae cranial CoM shift are evident in the first mass moments of most segments. However, the closest matches of whole-body CoM and these moments (large deviation starting around Maniraptora (nodes 11 and 12), reaching a maximum in basal Ornithurae (node 15)), are evident only for the head, tail, pectoral and pelvic limbs (Fig. 4). Yet, correlation analysis (Spearman's rank; see Methods Summary and Supplementary Tables 9–12 for details and results) supports a significant ($P < 0.05$) positive relationship between first mass moments and whole-body CoM position only for the pectoral ($P < 0.01$, $R = 0.67$) and pelvic ($P = 0.02$, $R = 0.58$) limbs. Furthermore, separate analysis of segment mass (Fig. 4, red) and segment CoM (Fig. 4, blue) indicates that, morphologically, the influence of the pelvic limb on whole-body CoM is largely a result of cranial evolutionary shifts of the segment CoM ($P = 0.06$, $R = 0.50$, same method) associated with expansion of the preacetabular ilium and the cnemial crest of the tibia, both of which add mass cranially to the thigh. In contrast, the influence of the pectoral limb on whole-body CoM is mainly due to increases in its mass ($P = 0.05$, $R = 0.51$).

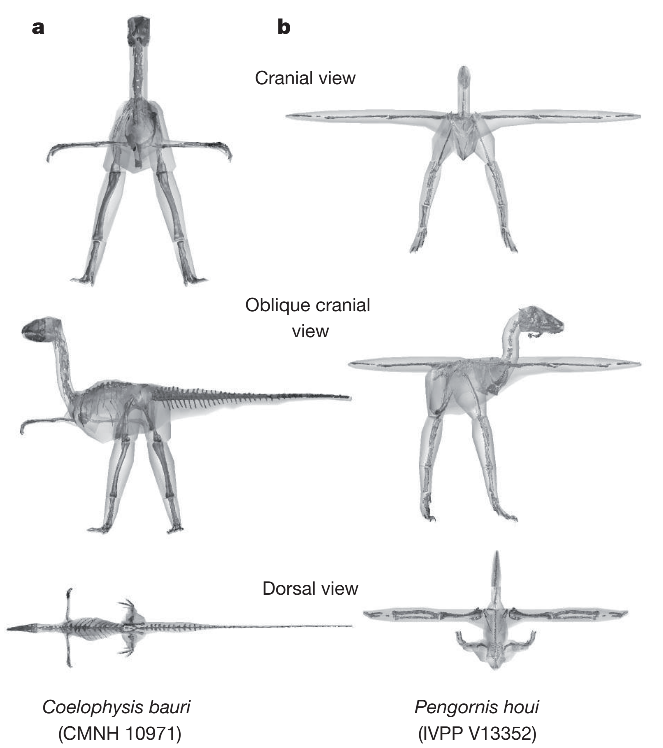


Figure 2 | Reconstructed body volumes. Based on digitized fossil skeletons and computed tomography scan data from modern relatives, for a basal dinosaur (**a**) and a basal bird (**b**); in cranial (top), oblique cranial (middle) and dorsal (bottom) views. These exemplify the major changes in body proportions that evolved on the bird line. Specimen numbers are shown under taxon names. For more details, see Supplementary Videos 1 and 2.

From the above findings, we infer that the Maniraptora-to-Ornithurae cranial CoM shift resulted from increased relative pectoral limb mass (Fig. 4d, red) and increasingly cranial segment CoMs for the pelvic limb (Fig. 4e, blue). Less significant ($P < 0.1$), but possibly important, positive correlations with a more cranial whole-body CoM are first mass moments for the head ($P = 0.08$, $R = 0.47$) and neck ($P = 0.08$, $R = 0.44$). On the basis of trends for these segments (Fig. 4a, b), we therefore suggest that, secondary to changes in limb morphology, a cranial shift in CoM may also have been associated with increased relative mass of the head and neck.

As predicted from gross anatomy[1,2], relative tail mass is estimated to have declined within Theropoda and tail CoM to have moved cranially (Fig. 4f, node 5), particularly within Maniraptora (node 11), to a minimum in basal Ornithurae (node 15). That the suggested[1,2] correlation between these trends and a more cranial whole-body CoM was not found to be significant (Supplementary Table 12) is notable. Considering that the tail represents the majority of body mass caudal to the hip, reduction or cranial concentration of tail mass, or both, would be expected to bias the whole-body CoM position strongly cranially. However, our results indicate that the effects of tail reduction were not significant in comparison to concurrent changes to the limbs (especially pectoral) and, to a lesser extent, the head and neck. Therefore we infer that adding mass to the front of theropod bodies was more influential for CoM evolution than was removing it from the back.

In addition to overall tail mass, we used volumetric reconstruction[21,22] to estimate evolutionary trends in the relative mass of the M. caudofemoralis longus (CFL) muscle (Fig. 5). The CFL is a principal locomotor muscle in most non-avian Reptilia, and was probably so in

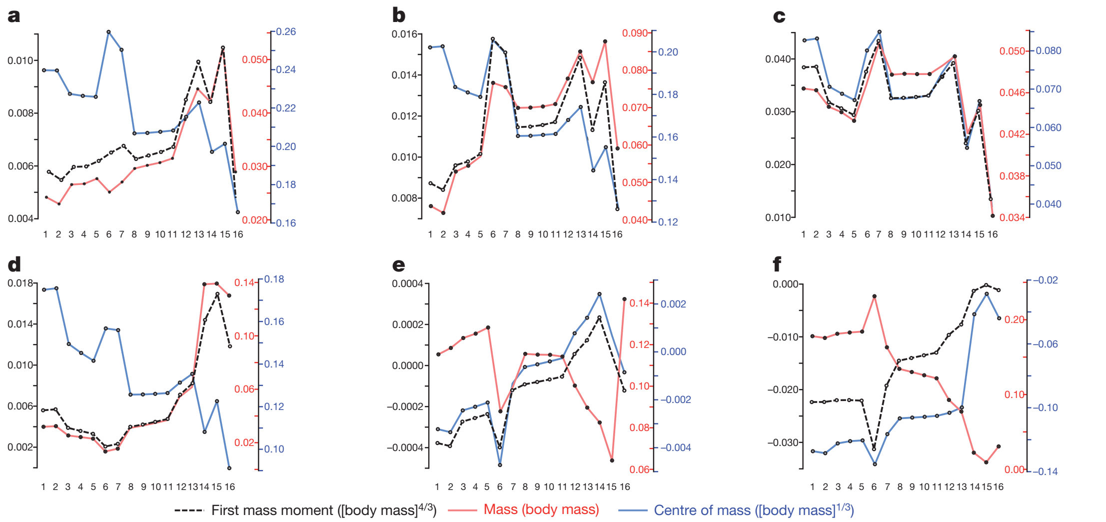


Figure 4 | Estimated evolutionary trends. Shown are trends in individual segment first mass moments (black dashed lines), individual segment masses (red lines) and individual segment craniocaudal CoM positions (blue lines) for the head (**a**), neck (**b**), trunk (**c**), pectoral limb (**d**), pelvic limb (**e**) and tail (**f**). Numbers 1–16 correspond to the nodes in Fig. 1. For reference, node 5 is Theropoda and node 13 is Aves.

ancestral archosaurs and dinosaurs as well[1,23]. It extends from the tail to the proximal femur and knee, and retracts the femur powerfully through a large arc during the stance phase. In Neornithes, the CFL has atrophied, and femoral retraction during walking is mostly replaced by knee flexion powered by enlarged 'hamstring' muscles[24,25]. Because the CFL's mass would have been a major locomotor power source in ancestral archosaurs, its mass is a reasonable proxy for its relative importance in hip extension or femoral retraction. Previously, tail reduction and the evolution of a suite of anatomical hip features associated with novel, long-axis control of the femur have been used to infer that the transition between tail-based to knee-based locomotion (and crouched limbs) began in earlier theropods[1–4]. Specifically, it was inferred that this trend began within basal Tetanurae, and that a derived system had already evolved in the clade Eumaniraptora[3].

Our estimates of CFL mass support some elements of this hypothesis, in that the CFL and tail mass are strongly reduced from Eumaniraptora onwards (Fig. 5). However, we estimate that the CFL muscle remained relatively large in basal Tetanurae despite overall tail mass reduction (Fig. 5, nodes 7–9, dotted line), indicating that locomotion remained plesiomorphically more hip driven than knee driven. In addition, our CoM and other body proportion estimates do not unambiguously support alterations of posture at these more basal nodes (Figs 3 and 4). Therefore, the origin of novel hip control features in basal Tetanurae may not have been directly associated with or driven by a postural shift, but

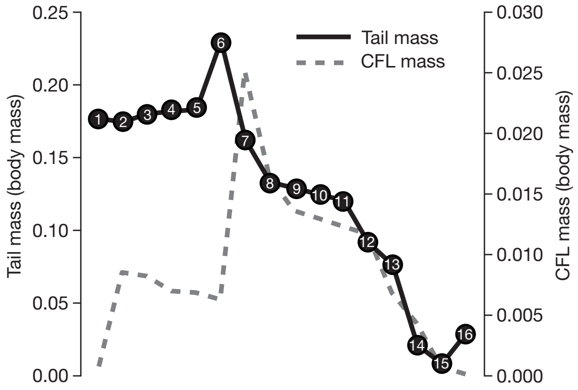


Figure 5 | Estimated evolutionary trends for tail segment and M. caudofemoralis longus muscle masses. Numbers 1–16 correspond to the nodes in Fig. 1.

instead may have been co-opted for later usage in supporting a more crouched posture.

Our results have clear implications for the evolution of bipedal locomotion along the bird line. The pattern of cranial CoM migration, proportional evolution and CFL reduction reconstructed here supports a gradual, stepwise acquisition of more-crouched limb postures across much of theropod evolution[1–4], rather than a rapid transition from more-upright postures occurring around the base of Avialae[5,6,8]. Our models explicitly yield the strongest support for a locomotor transition within the clade Maniraptora, and, perhaps more conservatively, Eumaniraptora (by which time the trend is well under way (Fig. 3, node 12)), in which considerable cranial CoM migration and concomitant strong reduction in CFL mass (Fig. 5, node 11 onwards) occurred. The fully derived modern condition probably did not evolve until well within Aves (for example Ornithurae (Figs 3 and 4, node 15)), when CoM position reached its cranial maximum and the CFL was most reduced[1–4]. Rather than being a phenomenon associated with or driven by tail reduction, we instead find that enlargement of the pectoral limb into the 'raptorial' forelimbs (and, ultimately, wings) of many eumaniraptorans is the strongest associated morphological trend. However, a more cranially biased pelvic limb CoM and perhaps increased head and neck mass were also involved. Note that this is also without considering the added mass of pectoral plumage (the geometry of which is too uncertain to model rigorously), particularly the large primary or primary-like feathers of Maniraptoriformes and later bird-line taxa (see, for example, ref. 26), which would only strengthen the relationship of cranially shifted body CoM and pectoral mass. Additional support for a locomotor transition within Eumaniraptora comes from the evolution of highly retroverted pubes, which, as previous studies have proposed, is likely to have fundamentally altered the moment arms (and, by inference, functions) of several major locomotor muscles[3,27,28].

Detailed phylogenetic and temporal aspects of the evolution of flight in the bird line remain controversial[4]. However, our finding that accelerated morpho-functional trends commenced around the node Eumaniraptora is closely correlated with the origin and diversification of animals with some degree of flight capability. Until more robust phylogenetic and aerodynamic assessments for early maniraptoriforms are made, it is impossible to assess conclusively whether our predictions of CoM and body shape change preceded, coincided with or followed the origin of flight. Our openly available data set (Methods Summary) and novel whole-body evolutionary approach mean that future studies can use our data to address these and other controversies. For example,

the addition of accurate feathering to our models of *Microraptor* and *Archaeopteryx* could result in a reassessment of the position of the centre of lift relative to that of the CoM (important for gliding and stable flight) or more complex flight aerodynamics[29]. However, our discovery that the evolution of CoM on the bird line was more influenced by body shape changes cranial to the hips than in the caudal region reverses the widely accepted view[1–8] and opens new questions about the degree of independence between fore- and hindlimb function[7] (that is, modularity) across this transition. The proposed relationship between novel hip control mechanisms and more-crouched pelvic limbs[3], and the linkage proposed here between pectoral limb size, CoM position and hindlimb posture, suggest that the evolution of both aerial and terrestrial locomotor anatomy were highly interconnected. Aerially adapted pectoral limbs and terrestrially adapted pelvic limbs belong to the same body, and the physical characteristics of one cannot logically be changed without affecting the mechanical functioning of the other. This reinforces the importance of whole-body biomechanical analysis in interpreting morpho-functional data from the fossil record.

METHODS SUMMARY

Body segment masses and CoM positions were estimated from computer reconstructions based on digitized skeletons. Fossil specimens were digitized (Supplementary Table 1; various scanners and settings; MIMICS 13 segmentation software, Materialise). Reconstructions of body shape were made in three-dimensional modelling software (BLENDER 2.49 (http://www.blender.org)); Autodesk MAYA 2012) using established methodology[18,22,28,30]. CoMs for individual body segments were analysed using custom code, and whole-body CoM was calculated using the equation

$$\mathrm{CoM}_{[x,y,z]} = \frac{1}{M}\sum_{i=1}^{n} m_i r_{i[x,y,z]} \qquad (1)$$

Here M is the total body mass, m_i is mass of segment i and r_i is the distance from system origin to the CoM of segment i (calculated separately for each set of x, y and z coordinates). The term $m_i r_i$ (first mass moment) represents the total influence of segment i on the overall system CoM.

Maximal and minimal iterations of body segments were made in steps of $\pm 20\%$ of the radial dimensions (adjusted for cross-sectional profile) away from our initial 'best estimate' models, on the basis of the minimum variation (about mean values) in the extra-skeletal dimensions of saurian tails[18]. This is probably too generous for less 'fleshy' segments; a more complete study of such dimensions is needed. Segment iterations were combined to represent the most cranial, caudal, dorsal and ventral distributions of mass and maximal and minimal overall mass (Supplementary Video 1). Mass properties were estimated using validated custom software[18,30]. Data and software code used are deposited in the Dryad repository at http://dx.doi.org/10.5061/dryad.hh74n.

CoM positions and segment masses were then normalized (divided by body mass or the cube root of body mass) and used to reconstruct ancestral node states with the 'trace characters' function (squared-change parsimony) in Mesquite 2.75 phylogenetic analysis software, using the phylogeny in Fig. 1 and estimated branch lengths in millions of years. See Supplementary Tables 3–12 for data sets and analysis results. Owing to non-normality, associations between normalized segment morphometrics (first mass moment, mass and CoM) were assessed using a non-parametric correlation test in R ('Hmisc' package, Spearman's rank).

Received 13 June 2012; accepted 7 March 2013.
Published online 24 April 2013.

1. Gatesy, S. M. Caudofemoral musculature and the evolution of theropod locomotion. *Paleobiology* **16,** 170–186 (1990).
2. Gatesy, S. M. in *Functional Morphology in Vertebrate Paleontology* (ed. Thomason, J. J.) Ch. 13 (Cambridge Univ. Press, 1995).
3. Hutchinson, J. R. & Gatesy, S. M. Adductors, abductors, and the evolution of archosaur locomotion. *Paleobiology* **26,** 734–751 (2000).
4. Hutchinson, J. R. & Allen, V. The evolutionary continuum of limb function from early theropods to birds. *Naturwissenschaften* **96,** 423–448 (2009).
5. Gatesy, S. M. Hind limb scaling in birds and other theropods: implications for terrestrial locomotion. *J. Morphol.* **209,** 83–96 (1991).
6. Christiansen, P. & Bonde, N. Limb proportions and avian terrestrial locomotion. *Geology* **371,** 356–371 (2002).
7. Gatesy, S. M. & Dial, K. P. Locomotor modules and the evolution of avian flight. *Evolution* **50,** 331–340 (1996).
8. Jones, T. D., Farlow, J. O., Ruben, J. A., Henderson, D. M. & Hillenius, W. J. Cursoriality in bipedal archosaurs. *Nature* **406,** 716–718 (2000).
9. Herr, H. & Popovic, M. Angular momentum in human walking. *J. Exp. Biol.* **211,** 467–481 (2008).
10. Roberts, T. J. & Scales, J. A. Mechanical power output during running accelerations in wild turkeys. *J. Exp. Biol.* **205,** 1485–1494 (2002).
11. Clark, J. & Alexander, R. M. Mechanics of running by quail (*Coturnix*). *J. Zool.* **176,** 87–113 (1975).
12. Hancock, J. A., Stevens, N. J. & Biknevicius, A. R. Whole-body mechanics and kinematics of terrestrial locomotion in the Elegant-crested Tinamou *Eudromia elegans*. *Ibis* **149,** 605–614 (2007).
13. Roberts, T. J., Chen, M. S. & Taylor, C. R. Energetics of bipedal running II: limb design and running mechanics. *J. Exp. Biol.* **2762,** 2753–2762 (1998).
14. Carrano, M. T. & Biewener, A. A. Experimental alteration of limb posture in the chicken (*Gallus gallus*) and its bearing on the use of birds as analogues for dinosaur locomotion. *J. Morphol.* **240,** 237–249 (1999).
15. Biewener, A. A., Farley, C. T., Roberts, T. J. & Temaner, M. Muscle mechanical advantage of human walking and running: implications for energy cost. *J. Appl. Physiol.* **97,** 2266–2274 (2004).
16. Hutchinson, J. R. Biomechanical modeling and sensitivity analysis of bipedal running ability. I. Extant taxa. *J. Morphol.* **262,** 421–440 (2004).
17. Tarsitano, S. Stance and gait in theropod dinosaurs. *Acta Palaeontol. Pol.* **28,** 251–264 (1983).
18. Allen, V., Paxton, H. & Hutchinson, J. R. Variation in center of mass estimates for extant sauropsids and its importance for reconstructing inertial properties of extinct archosaurs. *Anat. Rec.* **292,** 1442–1461 (2009).
19. Gatesy, S. M. & Middleton, K. M. Bipedalism, flight, and the evolution of theropod locomotor diversity. *J. Vertebr. Paleontol.* **17,** 308–329 (1997).
20. Farlow, J. O., Gatesy, S. M., Holtz, T. R. J., Hutchinson, J. R. & Robinson, J. M. Theropod locomotion. *Am. Zool.* **40,** 640–663 (2000).
21. Persons, W. S. & Currie, P. J. The tail of *Tyrannosaurus*: reassessing the size and locomotive importance of the M. caudofemoralis in non-avian theropods. *Anat. Rec.* **294,** 119–131 (2011).
22. Hutchinson, J. R., Bates, K. T., Molnar, J., Allen, V. & Makovicky, P. J. A Computational analysis of limb and body dimensions in *Tyrannosaurus rex* with implications for locomotion, ontogeny, and growth. *PLoS ONE* **6,** e26037 (2011).
23. Gatesy, S. M. An electromyographic analysis of hindlimb function in *Alligator* during terrestrial locomotion. *J. Morphol.* **234,** 197–212 (1997).
24. Gatesy, S. M. Guineafowl hind limb function. II: electromyographic analysis and motor pattern evolution. *J. Morphol.* **240,** 127–142 (1999).
25. Marsh, R. L., Ellerby, D. J., Henry, H. T. & Rubenson, J. The energetic costs of trunk and distal-limb loading during walking and running in guinea fowl *Numida meleagris* I. Organismal metabolism and biomechanics. *J. Exp. Biol.* **209,** 2050–2063 (2006).
26. Turner, A. H., Makovicky, P. J. & Norell, M. A. Feather quill knobs in the dinosaur *Velociraptor*. *Science* **317,** 1721 (2007).
27. Hutchinson, J. R. The evolution of hindlimb tendons and muscles on the line to crown-group birds. *Comp. Biochem. Physiol. A* **131,** 169–197 (2001).
28. Hutchinson, J. R., Ng-Thow-Hing, V. & Anderson, F. C. A 3D interactive method for estimating body segmental parameters in animals: application to the turning and running performance of *Tyrannosaurus rex*. *J. Theor. Biol.* **246,** 660–680 (2007).
29. Chatterjee, S. & Templin, R. J. Biplane wing planform and flight performance of the feathered dinosaur *Microraptor gui*. *Proc. Natl Acad. Sci. USA* **104,** 1576–1580 (2007).
30. Bates, K. T., Manning, P. L., Hodgetts, D. & Sellers, W. I. Estimating mass properties of dinosaurs using laser imaging and 3D computer modelling. *PLoS ONE* **4,** e4532 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the following people and institutions for access to fossil specimens: S. Chapman, A. Milner, P. Holroyd, M. Goodwin, K. Padian, M. Ryan, G. Jackson, D. Chapman, H.-J. Siber, B. Pabst, Z. Zhou and M. Norell; and the NHM (UK), UCMP (USA), CMNH (USA), IVPP (Canada); Sauriermuseum Aathal (Switzerland) and AMNH (USA). We wish to thank the following people, institutions and companies for providing digitized specimens, reconstructed specimens or both: O. Grillo, H. Mallison, J. Hertel, J. Brougham, M. Davis, J. A. Bannister; and the Universidade Federal do Rio de Janeiro (Brazil), MNB (Germany), Crescendo Games (Canada), NOVA/WGBH (USA) and Mechanimal (USA). We thank J. Molnar and RVC for invaluable assistance in processing computed tomography and laser-scan data, and for video editing. This work was supported by the following grants and institutions: NERC grant no. NE/G005877/1 to J.R.H., a Royal Society International Joint Project to J.R.H. and Z. Zhou (not a co-author), and the Sam and Doris Welles Fund (University of California) as part of PhD funding to V.A.

Author Contributions K.T.B., V.A. and Z.L. digitized fossil material. V.A. and K.T.B. constructed and analysed volumetric reconstructions. J.R.H. and V.A. performed phylogenetic optimization analysis. V.A. performed all statistical analyses. J.R.H. supervised and contributed ideas throughout the project. All authors contributed to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.R.H. (jrhutch@rvc.ac.uk).

LETTER

doi:10.1038/nature12065

Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA

Muhammed Murtaza[1]*, Sarah-Jane Dawson[1,2]*, Dana W. Y. Tsui[1]*, Davina Gale[1], Tim Forshew[1], Anna M. Piskorz[1], Christine Parkinson[1,2], Suet-Feung Chin[1], Zoya Kingsbury[3], Alvin S. C. Wong[4], Francesco Marass[1], Sean Humphray[3], James Hadfield[1], David Bentley[3], Tan Min Chin[4,5], James D. Brenton[1,2,6], Carlos Caldas[1,2,6] & Nitzan Rosenfeld[1]

Cancers acquire resistance to systemic treatment as a result of clonal evolution and selection[1,2]. Repeat biopsies to study genomic evolution as a result of therapy are difficult, invasive and may be confounded by intra-tumour heterogeneity[3,4]. Recent studies have shown that genomic alterations in solid cancers can be characterized by massively parallel sequencing of circulating cell-free tumour DNA released from cancer cells into plasma, representing a non-invasive liquid biopsy[5–7]. Here we report sequencing of cancer exomes in serial plasma samples to track genomic evolution of metastatic cancers in response to therapy. Six patients with advanced breast, ovarian and lung cancers were followed over 1–2 years. For each case, exome sequencing was performed on 2–5 plasma samples (19 in total) spanning multiple courses of treatment, at selected time points when the allele fraction of tumour mutations in plasma was high, allowing improved sensitivity. For two cases, synchronous biopsies were also analysed, confirming genome-wide representation of the tumour genome in plasma. Quantification of allele fractions in plasma identified increased representation of mutant alleles in association with emergence of therapy resistance. These included an activating mutation in *PIK3CA* (phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha) following treatment with paclitaxel[8]; a truncating mutation in *RB1* (retinoblastoma 1) following treatment with cisplatin[9]; a truncating mutation in *MED1* (mediator complex subunit 1) following treatment with tamoxifen and trastuzumab[10,11], and following subsequent treatment with lapatinib[12,13], a splicing mutation in *GAS6* (growth arrest-specific 6) in the same patient; and a resistance-conferring mutation in *EGFR* (epidermal growth factor receptor; T790M) following treatment with gefitinib[14]. These results establish proof of principle that exome-wide analysis of circulating tumour DNA could complement current invasive biopsy approaches to identify mutations associated with acquired drug resistance in advanced cancers. Serial analysis of cancer genomes in plasma constitutes a new paradigm for the study of clonal evolution in human cancers.

Serial sampling of the tumour genome is required to identify the mutational mechanisms underlying drug resistance[2]. Serial tumour biopsies are invasive and often unattainable. Tumours are heterogeneous and continuously evolve, and even if several biopsies are obtained, these are limited both spatially and temporally. Analysis of isolated circulating tumour cells (CTCs) has been proposed, but circulating tumour DNA (ctDNA) is more accessible and easier to process[15]. Previous studies of tumour mutations in plasma have analysed individual loci, genes or structural variants to quantify tumour burden and to detect previously-characterized resistance-conferring mutations[1,6,16–18]. Genome-wide sequencing of plasma samples is used in prenatal diagnostics, demonstrating comprehensive coverage of the genome[19]. More recently, genome-wide sequencing of plasma DNA has been demonstrated as a potential tool for detection of disease or analysis of tumour burden in patients with advanced cancers[5,7]. These studies established that plasma DNA contains representation of the entire tumour genome[7], mixing together variants originating from multiple independent tumours[5]. This suggests that deeper sequencing of plasma DNA, applied to selected samples with high tumour burden in blood, may allow assessment of clonal heterogeneity and selection. In this study, we applied exome sequencing of ctDNA as a platform for non-invasive analysis of tumour evolution during systemic cancer treatment (Fig. 1).

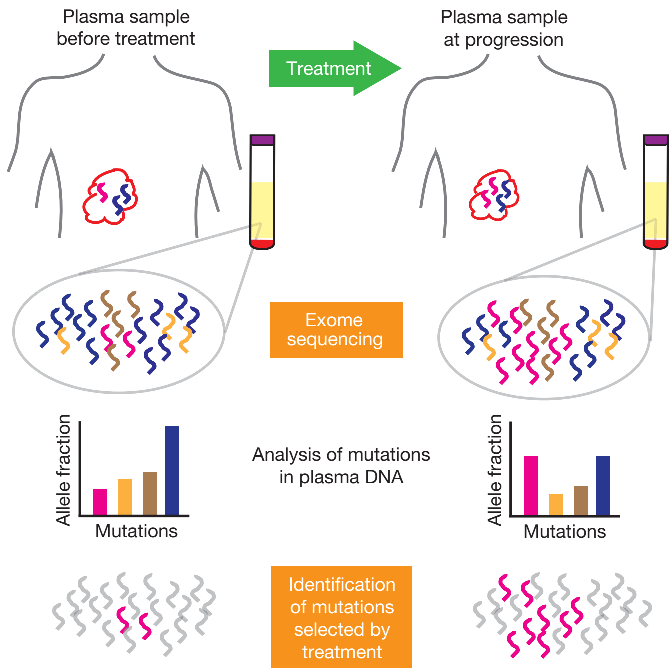


Figure 1 | Identification of treatment-associated mutational changes from exome sequencing of serial plasma samples. Overview of the study design: plasma was collected before treatment and at multiple time-points during treatment and follow-up of advanced cancer patients. Exome sequencing was performed on circulating DNA from plasma at selected time-points, separated by periods of treatment, and germline DNA. Mutations were identified across the plasma samples, and their abundance (allele fraction) at different time-points compared, generating lists of mutations that showed a significant increase in abundance, which may indicate underlying selection pressures associated with specific treatments. These lists contained mutations known to promote tumour growth and drug resistance, but also mutations of unknown significance. Accumulating such data across large cohorts could identify genes or pathways with recurrent mutations.

[1]Cancer Research UK Cambridge Institute and University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [2]Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. [3]Illumina, Inc., Chesterford Research Park, Little Chesterford CB10 1XL, UK. [4]Department of Haematology-Oncology, National University Cancer Institute, National University Health System, 5 Lower Kent Ridge Road, Tower block level 7, 119074 Singapore. [5]Cancer Science Institute, National University of Singapore, Centre for Translational Medicine, 14 Medical Drive, #12-01, 117599 Singapore. [6]Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK.
*These authors contributed equally to this work.

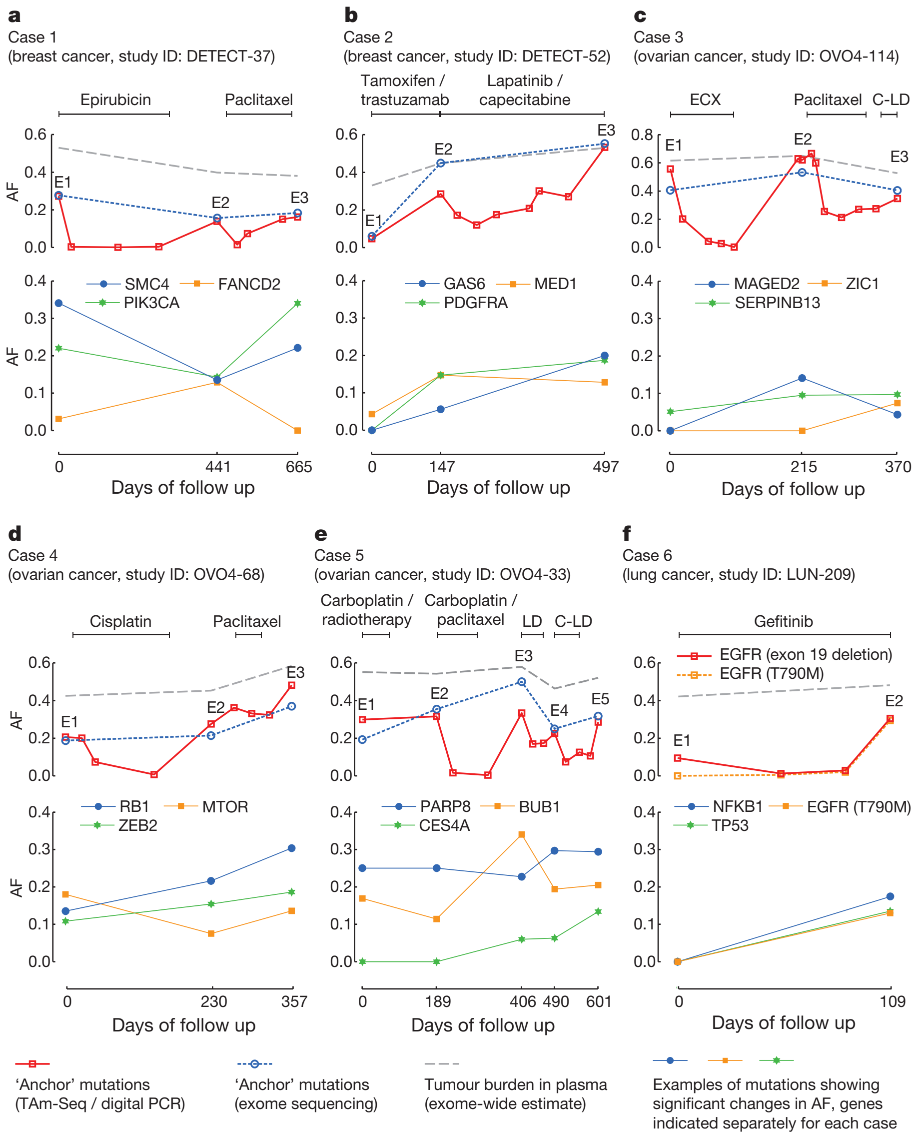


Figure 2 | Mutations showing evidence of genomic tumour evolution. All panels (**a**–**f**) are made up of an upper and a lower subpanel. Upper subpanels, time courses for allele fractions (AF; data points) of 'anchor' mutations used for initial quantification of ctDNA levels, and the fractional concentration of tumour DNA (tumour burden; grey dashed lines). 'Anchor' mutations were measured using digital PCR or TAm-Seq[6] for all available plasma samples, and using exome sequencing at selected time points indicated by E1, E2, E3 (and E4 and E5 for case 5). Tumour burden was estimated from exome data (an adaptation of genome-wide aggregated allelic loss[7]). In **a**, AF was averaged over six mutations measured in parallel using digital PCR. In **b**, a single mutation in *ATM* (predicted amino acid change I2948F) was measured by TAm-Seq. In **c**, **d** and **e**, a single mutation in *TP53* was measured by digital PCR for each case (R175H, K132N and R175H, respectively). In **f**, digital PCR was used to measure abundance of a deletion in exon 19 of *EGFR* (not quantified in exome sequencing data) and the EGFR T790M mutation. Lower subpanels, AF in exome data for selected mutations (blue, green and orange datapoints, see key) for each of the cases. Additional details are listed in Table 1, and a full list of mutations that showed a significant increase in abundance is included in Supplementary Tables 2–7. ECX, epirubicin, cisplatin and capecitabine; C-LD, carboplatin and liposomal doxorubicin; LD, liposomal doxorubicin.

We performed whole exome sequencing of plasma DNA in six patients with advanced cancers (Supplementary Table 1): two with breast cancer (cases 1 and 2), three with ovarian cancer (cases 3–5), and one with non-small-cell lung cancer (NSCLC, case 6). Exome sequencing was performed on multiple plasma samples from each patient separated by consecutive lines of therapy, spanning up to 665 days of clinical follow up (range 109–665 days, median 433 days). The ability to detect genomic events using redundant sequencing is dependent on the allele fraction (AF) of the mutant alleles in the samples analysed (ratio of mutant reads to depth of coverage at that locus), the sequencing depth, and the background noise rates of sequencing. Levels of ctDNA were previously quantified in these patients using digital PCR and tagged-amplicon deep sequencing[6] (TAm-Seq; Fig. 2, upper subpanels), allowing us to focus on samples with a high mutant AF in plasma, in which genomic changes related to the tumour could be identified even at relatively modest depth of sequencing. Comparison of AF measured using exome sequencing, digital PCR and TAm-Seq showed a high degree of concordance (correlation coefficient 0.8, $P < 0.0001$; Supplementary Fig. 1). Using as little as 2.3 ng of DNA (4%–20% of the DNA extracted from 2.0–2.2 ml of plasma), and an average of 169 million reads of sequencing per sample, we analysed the coding exons of all protein-coding genes at an average unique coverage depth ranging from 31-fold to 160-fold across 19 plasma samples (Supplementary Table 2). Consistent with previous reports[5,7], we observed copy number aberrations (CNAs, both gains and losses) in plasma samples in all patients across the whole genome (Supplementary Figs 2–7). These were strongly modulated by the fraction of tumour DNA in plasma and were particularly prominent in plasma samples in which mutant AF exceeded 50%.

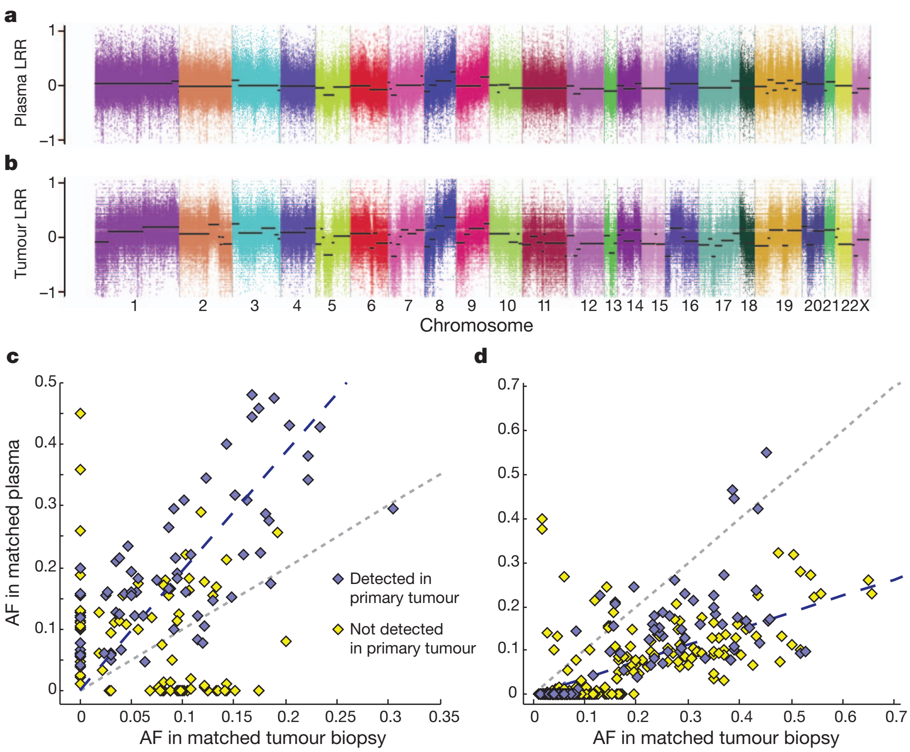


Figure 3 | Genome-wide concordance between plasma DNA and tumour DNA. **a**, **b**, Sequencing data were used to assess CNAs in the plasma sample (**a**) and in the synchronous metastatic tumour biopsy (**b**) from case 4. Panels show log *R* ratio (LRR), calculated on the basis of exome data, between plasma DNA and normal DNA (**a**) and between tumour and normal DNA (**b**). **c**, AF of mutations identified in exome data from plasma or metastatic biopsy for case 1. Grey dotted line shows equality. Blue dashed line has a slope of 1.93, indicating the median of the AF ratio for mutations found in both samples. Key applies to **c** and **d**. **d**, As **c** but for case 4, blue dashed line has a slope of 0.37.

For two cases, sequencing data were also available from metastatic tumour biopsies, collected at the same time as plasma samples (case 1 sample E1, and case 4 sample E2), and from tumour samples collected at the patients' initial presentation, 9 and 4.5 years earlier. CNAs were concordant between plasma and metastasis DNA in both patients (Fig. 3a, b, and Supplementary Fig. 7). Mutations identified in sequencing data[20–23] from the plasma or metastatic biopsy were compared (Supplementary Information). In case 1 with breast cancer, 151 mutations were identified in either the plasma or the synchronous biopsy. Of these, 93 mutations were found in both, and mutant AFs for these were higher in the plasma sample compared to the metastatic biopsy. The correlation coefficient of mutant AFs was positive (0.71) for mutations that were also found in the primary tumour, but negative (−0.22) for other mutations (Fig. 3c). In case 4 with ovarian cancer, 895 mutations were identified in either plasma or the tumour biopsy. For 172 mutations found in both, AFs were positively correlated (0.72) and were higher in the metastatic biopsy, which also contained 686 'private' mutations with AF $<$ 0.2 that were not found in either the plasma or the earlier tumour sample (Fig. 3d).

To identify changes in the mutation profiles of the tumours, we compared the abundance of somatic mutations found in plasma before and after each course of systemic treatment. For each patient, we examined a conservative list of mutations, including all mutations that were called in any of the plasma samples with a Bonferroni-corrected binomial probability of $<$0.05 assuming a background sequencing error rate of 0.1%. For each mutation and course of treatment (spanned by a pair of plasma samples), a *P*-value for a possible change in mutant AF was calculated as the binomial probability of obtaining the observed number of mutant reads, given the sequencing depth and the observed abundance in the paired time-point, normalized by the fractional concentration of tumour-derived DNA in the plasma (based on genome-wide aggregated allelic loss[5], Supplementary Table 3). Overall, 364 non-synonymous mutations passed with false discovery rate of $<$10% for significant changes in normalized abundance, ranging from 15 to 121 for each case (median 49). These include mutations in well-known cancer genes, genes linked to drug resistance and drug metabolism, and genes not previously associated with carcinogenesis or therapy resistance (Supplementary Tables 4–9). Selected examples are shown in Table 1 and Fig. 2.

We highlight here five examples. In case 1 with breast cancer, a strong increase was observed in the abundance of an activating mutation in *PIK3CA* following treatment with paclitaxel (Fig. 2a and Table 1). This mutation has been shown to promote resistance to paclitaxel in mammary epithelial cells[8]. In case 2, a patient with an oestrogen-receptor (ER)-positive, HER2-positive breast cancer, treatment with tamoxifen in combination with trastuzumab led to an increase in abundance of a nonsense mutation near the carboxy terminus of *MED1*, an ER co-activator that has been shown to be involved in tamoxifen resistance[10,11]. After further treatment of this patient with lapatinib in combination with capecitabine, we observed an increase in abundance of a splicing mutation in *GAS6*, the ligand for the tyrosine kinase receptor AXL (Fig. 2b, Table 1). Activation of the AXL kinase pathway has been shown to cause resistance to tyrosine kinase inhibitors in NSCLC[13] and resistance to lapatinib in ER-positive, HER2-positive breast cancer cell lines[12]. In case 4 with ovarian cancer, following treatment with cisplatin, we observed increase in abundance of a truncating mutation in the tumour-suppressor *RB1* (Fig. 2d, Table 1), predicted to inactivate the RB1 protein (Supplementary Fig. 8). In the matched metastasis biopsy obtained after treatment, the mutation was found in 95% of sequencing reads (59 of 62), with apparent loss of heterozygosity at 13q containing the *RB1* gene (Fig. 3a, b). Loss of *RB1* has been linked with chemotherapy response[9]. Case 6 was a NSCLC patient with an activating mutation in *EGFR* who was treated with gefitinib but progressed on treatment. Analysis by digital PCR detected the EGFR T790M mutation in plasma at progression, but not at the start of treatment. This mutation inhibits binding of

Table 1 | Selected mutations whose mutant AF significantly increased following treatment

Patient	Cancer type	Gene	Effect	Potential biological interest	Associated treatment	Mutant AF in plasma	
						Before	After
Case 1	Breast	*PIK3CA*	E545K	PI-3-kinase. p.E545K mutation associated with chemoresistance in mammary epithelial cells[8].	Paclitaxel	14%	34%
Case 1	Breast	*BMI1*	S324Y	BMI1 polycomb ring finger oncogene. Associated with chemoresistance[25].	Paclitaxel	3%	12%
Case 1	Breast	*SMC4*	I1000S	Structural maintenance of chromosomes 4. Downregulated in taxane resistant cell lines[26].	Paclitaxel	14%	22%
Case 1	Breast	*FANCD2*	G56V	Fanconi anaemia complementation group D2. Chromatin dynamics and DNA crosslink repair[27].	Epirubicin	3%	13%
Case 2	Breast	*MED1*	S1179X	Mediator complex subunit 1. Co-activator of ER with functional role in tamoxifen resistance[10,11].	Tamoxifen/trastuzumab	4%	15%
Case 2	Breast	*ATM*	I2948F	Ataxia telangiectasia mutated.	Tamoxifen/trastuzumab	6%	45%
Case 2	Breast	*PDGFRA*	D714E	Platelet-derived growth factor alpha. Cell surface tyrosine kinase receptor.	Tamoxifen/trastuzumab	0%	15%
Case 2	Breast	*GAS6*	Splicing	Growth arrest-specific 6. Ligand for AXL, overexpression associated with TKI resistance[12,13].	Lapatinib/capecitabine	6%	30%
Case 2	Breast	*TP63*	Splicing / S551G	Tumour protein p63.	Lapatinib/capecitabine	4%	20%
Case 4	Ovarian	*RB1*	E580X	Retinoblastoma 1. Loss of RB1 associated with EMT and drug resistance[9].	Cisplatin	14%	22%
Case 4	Ovarian	*ZEB2*	Y663C	Zinc finger E-box binding homeobox 2. Overexpression associated with cisplatin resistance in ovarian cancer[28].	Cisplatin	11%	15%
Case 4	Ovarian	*MTOR*	K1655N	Mechanistic target of rapamycin. Activating mutations in mTOR confers resistance to antimicrotubule agents[29].	Paclitaxel	8%	14%
Case 5	Ovarian	*CES4A*	P55S	Carboxylesterase 4A. Hydrolysis or transesterification of various xenobiotics.	Carboplatin/paclitaxel Carboplatin/liposomal doxorubicin	0% 6%	6% 13%
Case 5	Ovarian	*BUB1*	M889K	Mitotic checkpoint serine/threonine-protein kinase.	Carboplatin/paclitaxel	11%	34%
Case 5	Ovarian	*PARP8*	P81T	Poly [ADP-ribose] polymerase family, member 8.	Liposomal doxorubicin	23%	30%
Case 6	Lung	*EGFR*	T790M	Epidermal growth factor receptor. Established to cause gefitinib resistance by inhibiting drug binding[14].	Gefitinib	0%	13%
Case 6	Lung	*TP53*	Y163C	Tumour protein p53[30].	Gefitinib	0%	14%
Case 6	Lung	*NFKB1*	G489V	Nuclear factor κB[30].	Gefitinib	0%	17%

Potential biological role and associations with drug resistance described in literature are highlighted. The "Effect" column lists predicted change in amino acid sequence.

gefitinib to EGFR and has been established as the main driver of acquired resistance to gefitinib[14]. Unbiased analysis of plasma DNA by exome sequencing identified selection for this mutation amongst genomic changes that occurred following therapy (Fig. 2f, Table 1).

In this proof of principle study, we demonstrate that exome analysis of plasma ctDNA represents a novel paradigm for non-invasive characterization of tumour evolution. Our data, together with recent reports[5,7], show that CNAs and somatic mutations identified in ctDNA are widely representative of the tumour genome and provide an alternative method of tumour sampling that can overcome limitations of repeated biopsies. Cell-free DNA fragments from multiple lesions in the same individual all mix together in the peripheral blood[5], therefore ctDNA is likely to contain a wider representation of the genomes from multiple metastatic sites, whereas mutations present in a single biopsy or minor sub-clone may be missed. This strengthens the case for the use of ctDNA as a biomarker for monitoring tumour burden or for the analysis of hotspot mutation regions[1,6,16,17], but also indicates that tracking different mutations for assessment of tumour heterogeneity and clonal evolution is now possible. Our data identified a subset of genes that were positively selected following treatment, many of which have been previously associated with drug resistance. Other changes may represent 'passenger' mutations or false-positives, but some are likely to contribute to resistance to therapy. Accumulating data across a large number of cases could identify new genes or pathways that are frequently mutated following specific treatment types, and help refine analysis algorithms.

The approach we describe here may be broadly applicable to a large fraction of advanced cancers, where the median mutation burden in plasma (before start of treatment) is 5%–10% (refs 6, 16, 24). Analysis of acquired drug resistance is of particular utility in advanced or metastatic cancers, which is the target population for nearly all early phase clinical trials. Improvements in sequencing and associated technologies may enable similar analysis in cases with a lower tumour burden in plasma. At present, this non-invasive approach for characterizing cancer exomes in plasma is readily applicable to patients with high systemic tumour burden, enabling detailed and comprehensive evaluation of clonal genomic evolution associated with treatment response and resistance.

METHODS SUMMARY

Patients and samples. Cases 1–5 were recruited as part of prospective clinical studies at Addenbrooke's Hospital, Cambridge, UK, approved by the local research ethics committee (REC reference nos 07/Q0106/63, 08/H0306/61 and 07/Q0106/63). Case 6 was recruited as part of the 'Hydroxychloroquine and gefitinib to treat lung cancer' study (NCT00809237) at the National University Health System, Singapore, approved by the National Healthcare Group NHG IRB—DSRB 2008/00196. Written informed consent was obtained from patients, and serial blood samples were collected at intervals of ≥3 weeks.

Extraction and sequencing of plasma DNA. DNA was extracted from plasma using the QIAamp circulating nucleic acid kit (Qiagen) according to the manufacturer's instructions. Barcoded sequencing libraries were prepared using a commercially available kit (ThruPLEX-FD, Rubicon Genomics). Pooled libraries were enriched for the exome using hybridization (TruSeq Exome Enrichment Kit, Illumina), quantified using quantitative PCR and pooled in 1:1 ratio for paired-end sequencing on a HiSeq2500 (Illumina).

Variant calling and analysis. Sequencing data were demultiplexed and aligned to the hg19 genome using BWA[20]. Pileup files for properly paired reads with mapping quality ≥60 were generated using samtools[22]. AFs were calculated for all Q30 bases. A mutation was called if ≥4 mutant reads were found in plasma with ≥1 read on each strand, and no mutant reads were observed in germline DNA or in a prior plasma sample with ≥10-fold coverage. For comparison between consecutive plasma samples in a patient, we calculated the binomial probability of obtaining the observed AF (or greater) if the abundance of the mutant allele, normalized by tumour load in plasma (based on a modified genome-wide aggregated allelic loss method[5]), had remained constant between the two samples.

Full Methods and any associated references are available in the online version of the paper.

Received 5 October 2012; accepted 11 March 2013.
Published online 7 April 2013.

1. Diaz, L. A. Jr *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486,** 537–540 (2012).

2. Aparicio, S. & Caldas, C. The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368,** 842–851 (2013).
3. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366,** 883–892 (2012).
4. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).
5. Chan, K. C. *et al.* Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin. Chem.* **59,** 211–224 (2013).
6. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4,** 136ra168 (2012).
7. Leary, R. J. *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci. Transl. Med.* **4,** 162ra154 (2012).
8. Isakoff, S. J. *et al.* Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells. *Cancer Res.* **65,** 10992–11000 (2005).
9. Knudsen, E. S. & Knudsen, K. E. Tailoring to RB: tumour suppressor status and therapeutic response. *Nature Rev. Cancer* **8,** 714–724 (2008).
10. Cui, J. *et al.* Cross-talk between HER2 and MED1 regulates tamoxifen resistance of human breast cancer cells. *Cancer Res.* **72,** 5625–5634 (2012).
11. Nagalingam, A. *et al.* Med1 plays a critical role in the development of tamoxifen resistance. *Carcinogenesis* **33,** 918–930 (2012).
12. Liu, L. *et al.* Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. *Cancer Res.* **69,** 6871–6878 (2009).
13. Zhang, Z. *et al.* Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nature Genet.* **44,** 852–860 (2012).
14. Pao, W. *et al.* Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med.* **2,** e73 (2005).
15. Punnoose, E. A. *et al.* Evaluation of circulating tumor cells and circulating tumor DNA in non-small cell lung cancer: association with clinical endpoints in a phase II clinical trial of pertuzumab and erlotinib. *Clin. Cancer Res.* **18,** 2391–2401 (2012).
16. Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. *Nature Med.* **14,** 985–990 (2008).
17. McBride, D. J. *et al.* Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes Chromosom. Cancer* **49,** 1062–1069 (2010).
18. Yung, T. K. *et al.* Single-molecule detection of epidermal growth factor receptor mutations in plasma by microfluidics digital PCR in non-small cell lung cancer patients. *Clin. Cancer Res.* **15,** 2076–2084 (2009).
19. Lo, Y. M. *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2,** 61ra91 (2010).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
21. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
22. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
23. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
24. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl Acad. Sci. USA* **102,** 16368–16373 (2005).
25. Siddique, H. R. & Saleem, M. Role of BMI1, a stem cell factor, in cancer recurrence and chemoresistance: preclinical and clinical evidences. *Stem Cells* **30,** 372–378 (2012).
26. Chang, H. *et al.* Identification of genes associated with chemosensitivity to SAHA/taxane combination treatment in taxane-resistant breast cancer cells. *Breast Cancer Res. Treat.* **125,** 55–63 (2011).
27. Sato, K. *et al.* Histone chaperone activity of Fanconi anemia proteins, FANCD2 and FANCI, is required for DNA crosslink repair. *EMBO J.* **31,** 3524–3536 (2012).
28. Haslehurst, A. M. *et al.* EMT transcription factors snail and slug directly contribute to cisplatin resistance in ovarian cancer. *BMC Cancer* **12,** 91 (2012).
29. VanderWeele, D. J., Zhou, R. & Rudin, C. M. Akt up-regulation increases resistance to microtubule-directed chemotherapeutic agents through mammalian target of rapamycin. *Mol. Cancer Ther.* **3,** 1605–1613 (2004).
30. Wu, C. C., Yu, C. T., Chang, G. C., Lai, J. M. & Hsu, S. L. Aurora-A promotes gefitinib resistance via a NF-κB signaling pathway in p53 knockdown lung cancer cells. *Biochem. Biophys. Res. Commun.* **405,** 168–172 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Langmore and K. Solomon (Rubicon Genomics) for early access to library preparation products. We thank L. Jones, S. Richardson, C. Hodgkin and H. Biggs for recruiting patients into the DETECT and CTCR-OVO4 studies, all medical and ancillary staff in the breast and gynaecological cancer clinic and patients for consenting to participate. We thank the Human Research Tissue Bank at Addenbrooke's Hospital which is supported by the NIHR Cambridge Biomedical Research Centre. We thank the Cancer Science Institute, National University of Singapore, and the Hematology-Oncology Research Group, National University Health System, Singapore for their support. We acknowledge the support of Cancer Research UK, the University of Cambridge, National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge Experimental Cancer Medicine Centre, Hutchison Whampoa Limited, and the National Medical Research Council, Singapore. S.-J.D. is supported by an Australian NHMRC/RG Menzies Early Career Fellowship that is administered through the Peter MacCallum Cancer Centre, Victoria, Australia.

Author Contributions M.M., S.-J.D., T.F., D.W.Y.T., D.G., J.D.B., C.C. and N.R. designed the study. M.M., D.W.Y.T. and T.F. developed methods. S.-J.D., C.P., A.S.C.W., T.M.C., J.D.B. and C.C. designed and conducted the prospective clinical studies. M.M., S.-J.D., D.W.Y.T., D.G., T.F. and A.M.P. generated data. Z.K., S.H. and D.B. contributed sequencing data. M.M., F.M. and N.R. analysed sequencing data. S.-F.C. and J.H. contributed to experiments and data analysis. M.M., S.-J.D., D.W.Y.T., T.M.C., J.D.B., C.C. and N.R. interpreted data. M.M. and N.R. wrote the paper with assistance from S.-J.D., D.W.Y.T., C.C., J.D.B. and other authors. All authors approved the final manuscript. J.D.B., C.C. and N.R. are the project co-leaders and joint senior authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.D.B. (james.brenton@cruk.cam.ac.uk), C.C. (carlos.caldas@cruk.cam.ac.uk) or N.R. (nitzan.rosenfeld@cruk.cam.ac.uk).

METHODS

Sample collection. Cases 1–5: patients were recruited as part of prospective clinical studies at Addenbrooke's Hospital, Cambridge, UK, approved by local research ethics committee (REC reference nos 07/Q0106/63, 08/H0306/61 and 07/Q0106/63). Written informed consent was obtained from the patients. Serial blood samples were collected in EDTA tubes at intervals of ≥3 weeks, and centrifuged within 1 h at 820*g* for 10 min to separate the plasma from the peripheral blood cells. The plasma was then further centrifuged at 20,000*g* for 10 min to pellet any remaining cells. The plasma was then stored at −80°C until DNA extraction.

Case 6: this patient was recruited as part of the 'Hydroxychloroquine and gefitinib to treat lung cancer' study (NCT00809237) at the National University Health System, Singapore, approved by the National Healthcare Group NHG IRB-DSRB 2008/00196. Blood was collected in CPT tubes (BD Vacutainer) before gefitinib was started, and at monthly intervals while the patient was on treatment, until disease progression. Blood collected was spun within 1 h at 1,500*g* for 20 min, and the plasma fraction was frozen at −80 °C. Thawed samples were recentrifuged at 20,000*g* for 10 min to further separate any cellular portions.

Extraction of plasma DNA. DNA was extracted from aliquots of plasma using the QIAamp circulating nucleic acid kit (Qiagen) according to the manufacturer's instructions (see Supplementary Table 1 for volumes used). DNA was eluted into buffer AVE, eluted twice through each column to maximize yield, and stored at −20 °C.

Extraction of normal and tumour DNA. DNA from tumour sections was extracted using DNeasy tissue or DNA Allprep kits (Qiagen) according to manufacturer's instructions. Matched germline DNA was derived from normal peripheral blood leucocytes. After the collection of plasma from each blood sample, the remaining layer of normal peripheral blood lymphocytes ('buffy coat') was removed. This layer was either subjected to red cell lysis using a red cell lysis buffer (155 mM NH_4Cl, 10 mM $KHCO_3$ and 0.1 mM EDTA pH 7.4) and DNA extracted using a standard phenol-chloroform extraction protocol; or frozen at −80 °C before extraction using QIAamp DNA mini kit (Qiagen).

Sequencing of plasma DNA. Concentration of DNA for each plasma sample was determined using digital PCR, with an assay targeting *RPP30* for case 2, *TP53* for cases 3–5 and *EGFR* for case 6. For case 1, DNA concentration and 'anchor' mutation AF were calculated by averaging results from six assays targeting *PIK3CA*, *MET*, *IQCA1*, *CD1A*, *KIAA0406* and *ZFYVE21*. Libraries were generated using a commercially available kit for fragmented DNA (ThruPLEX-FD, Rubicon Genomics). 2.3–40 ng of DNA (Supplementary Table 2) was used to generate a sequencing library using manufacturer's protocols. Separate unique molecular identifiers were used for each sample. 30 µl of the library volume was obtained for each sample. 2–5 plasma DNA libraries from each patient were made and pooled together for exome capture using hybridization (TruSeq Exome Enrichment Kit, Illumina). Pools were concentrated using vacuum (Eppendorf Vacuum Concentrator) and prepared to 40 µl volume. Exome enrichment was performed following manufacturer's protocols. Enriched libraries were quantified using quantitative PCR and pooled in 1:1 ratio for paired-end next generation sequencing on HiSeq2500 (Illumina).

Sequencing of normal and tumour DNA. Sequence data for tumour and germline samples for case 1 have been reported previously. In brief, genomic libraries from tumour and matched normal tissue were prepared using the standard Illumina paired-end sample preparation kit according to the manufacturer's instructions. DNA fragments of 300 bp in size were sequenced using paired-end 100 bp reads on a HiSeq2000 (Illumina) achieving a depth of >30×. Germline samples for cases 2–6 and tumour sample for case 4 were sheared using Covaris and exome sequenced as described above.

Digital PCR. The principle of microfluidic digital PCR and its use for quantification of tumour DNA has been described previously[6,18]. Assays were designed based on TaqMan chemistry. All digital PCR analysis was carried out on the BioMark system using 12.765 Digital Arrays (Fluidigm) following manufacturer's instructions and protocol. Briefly, 3.5 µl from the eluted DNA was heated to 95 °C for 1 min and placed on ice, then mixed with TaqMan Universal PCR Master Mix (Applied Biosystems) and sample loading buffer (Fluidigm) into a final reaction volume of 10 µl and loaded into each panel of the chip. The reaction mix was then automatically partitioned into 765 reaction chambers. The numbers of starting template DNA molecules were calculated using Poisson statistics based on the number of positive amplifications[6,18].

Analysis of sequencing data. Sequencing reads were demultiplexed allowing zero mismatches in barcodes. Paired-end alignment to the hg19 genome was performed using BWA version 0.5.9 for all exome sequencing data including germline samples, plasma samples and tumour metastasis where generated[20]. PCR duplicates were marked using Picard. Local realignment was performed using Genome Analysis Tool Kit (GATK)[21]. Pileup files were generated for the genomic regions targeted by exome enrichment using samtools v0.1.17[22]. For plasma samples, properly paired reads with mapping quality ≥60 were used to generate the pileup. AFs for each single-base locus were calculated for all bases with phred quality ≥30. For germline DNA, an additional pileup file was generated (using a mapping quality cut-off of ≥1 and without any base quality cut-offs) and was used as reference for calling somatic variants. A mutation was called if no mutant reads for an allele were observed in germline DNA at a locus that was covered at least 10 fold, and if at least 4 reads supporting the mutant were found in the plasma data with at least 1 read on each strand (forward and reverse). At loci with <10-fold coverage in normal DNA and no mutant reads, mutations were called in plasma if a prior plasma sample showed no evidence of a mutation and was covered adequately (10 fold or more). All mutations were annotated for genes and function as well as repeated genomic regions using ANNOVAR[23].

AF was defined as the number of high quality reads supporting a mutation as a fraction of the total number of high quality reads covering the locus. For each patient, AF and number of reads for any mutations called with the above parameters were identified in all plasma samples. A binomial probability of obtaining the observed number of reads given depth in each plasma sample was calculated. The minimum of these probability values was corrected using Bonferroni correction for 62 million $\times$ n hypotheses tested, where n was the number of plasma samples sequenced (3 samples for cases 1–4, 5 samples for case 5 and 2 samples for case 6). Mutations with corrected P-values under 0.05 were retained for further analysis in plasma samples.

Estimation of CNAs. To assess CNAs, plasma DNA and tumour sequencing data were compared to germline DNA data at single nucleotide polymorphisms (SNPs) covered within the targeted exome region. The SNPs were identified from the publicly available 1000 Genomes Project data.

Depth information was normalized by dividing the depth of each SNP by the median depth across all SNPs. The log R ratio (LRR) was computed as the base-10 logarithm of the sample depth (metastasis or plasma) divided by the depth of the normal. Each chromosome was segmented by an iterative process that considered non-overlapping blocks of 1,000 data points. Points lying at least 1.5 standard deviations away from the median LRR for the block were removed from the mean LRR computation. If the difference in mean LRR between two consecutive blocks was less than 0.12, the blocks were merged into a single segment whose mean LRR was re-computed using points from both blocks.

Segmentation of B allele frequency (BAF) plots was similarly performed, considering windows of 1,000 data points and starting new segments if the difference in median frequency was greater than 4%. Blocks whose median frequency was within 8% of the median chromosome frequency in the normal sample were considered consistent with the BAF of the normal sample.

Comparison of mutations between plasma and tumour. For tumour/plasma comparison presented for cases 1 and 4, we identified all mutations called in data from synchronous plasma and metastatic tumour samples, as described above. We retained all mutations adequately covered in both samples (minimum 50 reads in plasma, minimum 10 reads in synchronous tumour whole genome data for case 1, minimum 50 reads in synchronous tumour exome data for case 4). We further discarded all mutations with no coverage in archived tumour samples obtained earlier (9 years earlier for case 1, and 4.5 years earlier for case 4).

Identification of mutations that changed in representation over treatment. To estimate systemic tumour burden, we calculated fractional concentration of ctDNA in blood using an adaptation of genome-wide aggregated allelic loss[5]. AFs of SNPs from the 1000 Genomes Project were obtained for germline and plasma data. SNPs with $0 < AF < 1$ in germline DNA were identified. SNPs where the minor AF in the germline data deviated from heterozygosity were identified using a binomial probability of obtaining the observed number of minor allele reads given depth in germline DNA and expected AF of 0.5. SNPs with probability <0.25 were discarded from further analysis.

Of the remaining SNPs, significant deviation from heterozygosity in any of the sequenced plasma samples, determined by a binomial distribution using sequencing depth and expected AF of 0.5, was used to identify loss of heterozygosity (LOH). SNPs with a probability <0.01 in any of the sequenced plasma samples were retained for estimation of tumour burden as described previously[5]. Fractional ctDNA burden was calculated as follows:

1 – [(sum of reads in the lost alleles)/(sum of reads in the retained alleles)]

AFs for all mutations were normalized by the estimated tumour burden. For any comparison between two consecutive plasma samples in a patient, we calculated the binomial probability for the observed difference in AF assuming no difference in normalized abundance. For a comparison between (for example) E1 and E2, we calculated the probability of obtaining the observed number of mutant reads or greater in E2 if normalized abundance in E2 had remained the same as in E1; this probability was multiplied by the probability of the observed number of mutant reads or less in E1 if the normalized abundance in E1 was the same as observed in E2. Where no mutant reads were obtained in the E1, only the reverse direction was used for this analysis. Changes in representation with a false discovery rate of 10% or lower, which were exonic non-synonymous or splicing mutations, were retained and are presented in Supplementary Tables 2–7.

doi:10.1038/nature12063

Random convergence of olfactory inputs in the *Drosophila* mushroom body

Sophie J. C. Caron[1], Vanessa Ruta[2], L. F. Abbott[1,3] & Richard Axel[1,4,5]

The mushroom body in the fruitfly *Drosophila melanogaster* is an associative brain centre that translates odour representations into learned behavioural responses[1]. Kenyon cells, the intrinsic neurons of the mushroom body, integrate input from olfactory glomeruli to encode odours as sparse distributed patterns of neural activity[2,3]. We have developed anatomic tracing techniques to identify the glomerular origin of the inputs that converge onto 200 individual Kenyon cells. Here we show that each Kenyon cell integrates input from a different and apparently random combination of glomeruli. The glomerular inputs to individual Kenyon cells show no discernible organization with respect to their odour tuning, anatomic features or developmental origins. Moreover, different classes of Kenyon cells do not seem to preferentially integrate inputs from specific combinations of glomeruli. This organization of glomerular connections to the mushroom body could allow the fly to contextualize novel sensory experiences, a feature consistent with the role of this brain centre in mediating learned olfactory associations and behaviours.

Olfactory perception in the fly is initiated by the binding of an odorant to an ensemble of olfactory sensory neurons (OSNs) in the antennae, resulting in the activation of a unique and topographically fixed combination of glomeruli in the antennal lobe (AL)[4,5]. The discrimination of odours therefore requires the integration of information from multiple glomeruli in higher olfactory centres. AL projection neurons (PNs) extend dendrites into a single glomerulus and project axons that bifurcate to innervate two distinct brain regions, the lateral horn and the mushroom body (MB)[6,7]. The invariant circuitry of the lateral horn is thought to mediate innate behaviours[8,9], whereas the MB translates olfactory sensory information into learned behavioural responses[1]. PN axons that innervate the MB terminate in large boutons[6,7] that synapse on Kenyon cells (KCs)[10–12]. A given KC extends a small number of dendritic "claws", with each claw receiving information from only one PN bouton[10–12]. A single bouton connects to multiple KC claws to form a discrete anatomic structure, the microglomerulus[10–12]. Each KC projects an axon to one of the three different classes of MB lobes, α/β, α'/β' or γ, where it synapses upon a relatively small number of extrinsic output neurons[13,14].

Electrophysiological and optical imaging studies show that odorants activate sparse subpopulations of KCs[2] distributed across the MB without spatial preference[3]. Individual KCs could be connected to preferential combinations of glomeruli that are co-ordinately activated by behaviourally relevant odours. Alternatively, KCs may not receive structured input; rather the glomerular inputs may be random, a feature that maximizes the diversity of odour representations in the MB. We have exploited the specialized structure of the PN–KC synapse to characterize the glomerular origin of the PNs that converge onto individual KCs.

Photoactivatable green fluorescent protein (PA-GFP) was expressed in all neurons of the fly and a single KC was photolabelled. We observe that individual photolabelled KCs elaborate between 2 and 11 dendritic claws (average = 7, $n = 200$) restricted to the main olfactory calyx (Fig. 1b, h and Supplementary Table 1). The axonal projections of a labelled KC can be traced into either the α/β, α'/β' or γ lobes of the MB (Fig. 1g and Supplementary Table 1). Texas red dextran was then electroporated into the centre of a single KC claw, filling the PN bouton innervating that claw (Fig. 1a–f). Retrograde transfer of the dye labels a single PN and its associated AL glomerulus ($n = 665$, Fig. 1g and Supplementary Table 1), providing further evidence that an individual KC claw receives input from only a single glomerulus.

We verified that this tracing method identifies functional connections between PNs and KCs. Functional imaging was performed on flies that express the calcium indicator protein GCaMP3 in most KCs to identify the claws activated by the stimulation of a single glomerulus (Fig. 2b). Electroporation of dye into an activated microglomerulus labels a single PN that innervates the stimulated glomerulus ($n = 10$, Fig. 2c–e). Thus, the electroporation of dye into a KC claw allows us to faithfully identify the PN to which it is functionally connected.

We used the strategy of photolabelling a single KC and sequential electroporation of dye into each of its claws to define glomerular inputs to an individual KC. In initial experiments, PA-GFP was expressed in all neurons and 100 randomly chosen KCs were photolabelled in 100 different flies. Among the 100 photolabelled KCs, 84 α/β KCs, 14 α'/β' KCs, but only 2 γ KCs were identified (Supplementary Table 1). Each MB contains about 1,000 α/β KCs, 370 α'/β' KCs and 670 γ KCs[15]. γ KCs are underrepresented in this initial data set. This is likely to result from the spatial segregation of their cell bodies, which renders γ KCs less accessible to photolabelling. Most γ KCs, but not the α/β and α'/β' KCs, express Fruitless (Fru)[16–18]. An additional 100 γ KCs were targeted for photoactivation in flies expressing PA-GFP under the control of the Fru promoter (Supplementary Table 1).

Texas red dextran was sequentially electroporated into different claws of a photolabelled KC (Fig. 1h, i). It is technically difficult to fill all the claws of a KC and on average three glomerular inputs were identified per KC (Fig. 3 and Supplementary Table 1). In fewer than 5% of the samples, the number of labelled PNs differed from the number of claws filled, reflecting either unsuccessful or imprecise electroporation. Samples with more labelled PNs than expected were discarded. The low frequency of unsuccessful claw fills indicates that claws extending from a given KC were filled with equal efficiency independently of size. Thus the size of a claw was not a selection criterion in these experiments.

A total of 683 inputs that synapse on 200 KCs were identified (Fig. 3 and Supplementary Table 1). We found that 654 of these inputs connect to PNs innervating 49 of the 51 AL glomeruli. PNs innervating the DA3 and VL1 glomeruli are absent from our data set, but we observe boutons from these PNs in the MB calyx (Supplementary Fig. 1). 29 of the claws receive input from brain regions other than the AL. Interestingly, 11 of these claws are innervated by PNs that derive from pseudoglomeruli in the proximal antennal protocerebrum, a thermosensing centre in the fly brain that receives input from distinct heat- and cold-sensing

[1]Department of Neuroscience, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA. [2]Laboratory of Neurophysiology and Behavior, The Rockefeller University, New York, New York 10065, USA. [3]Department of Physiology and Cellular Biophysics, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA. [4]Department of Biochemistry and Molecular Biophysics, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA. [5]Howard Hughes Medical Institute, Columbia University, New York, New York 10032, USA.

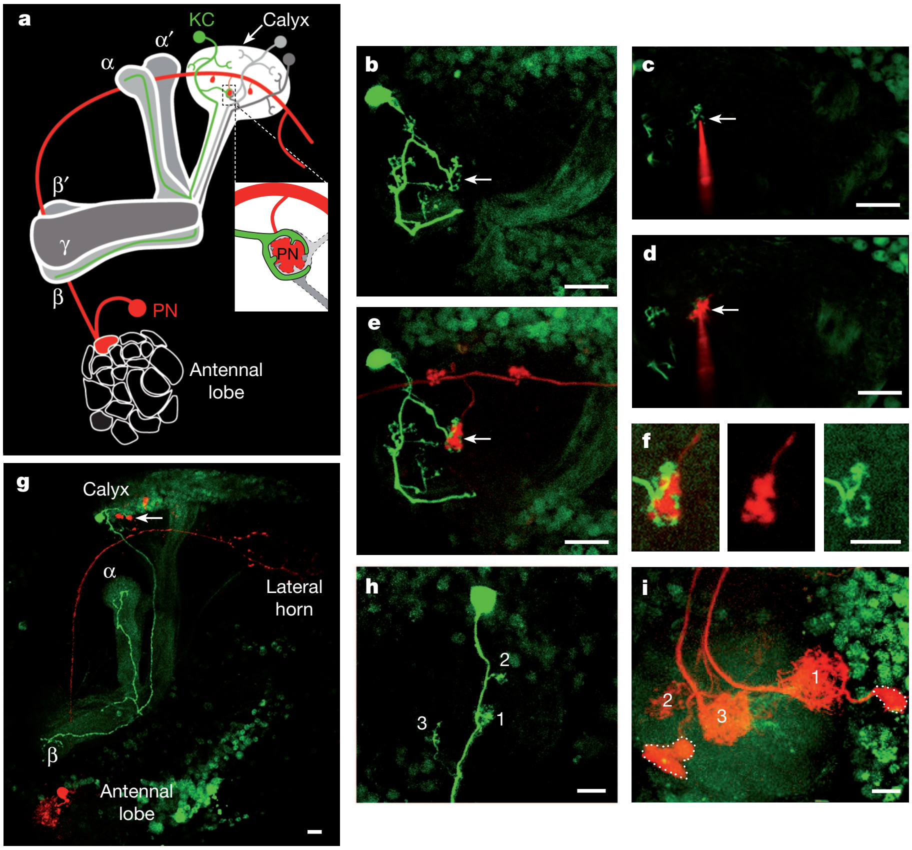


Figure 1 | Dye electroporation labels the PN connected to a KC claw. **a**, Schematic illustration of the tracing strategy used to identify the PN connected to a single KC claw. PNs (one shown in red) transmit olfactory information from a single glomerulus in the AL to the MB by forming multiple axonal boutons in the calyx. KCs extend dendrites into the MB calyx (white) and project axons into either the α/β (light grey), α′/β′ (medium grey), or γ (dark grey) lobe. The microglomerulus highlighted by a single photolabelled KC (green) is targeted for electroporation of red dye, resulting in the uptake of dye by a single PN and its associated AL glomerulus (red). Insert shows the targeted microglomerulus formed from a single red PN bouton connected to the photolabelled KC claw (green) as well as other unlabelled KC claws (different shades of grey). **b**, Photolabelling of a single KC expressing PA-GFP under the control of the pan-neuronal promoter synaptobrevin^{GAL4} reveals six dendritic claws within the MB calyx. **c**, An electrode filled with Texas Red dextran is centred into the microglomerulus outlined by one of the photolabelled KC claws shown in **b** (arrow). **d**, Dye is electroporated into the targeted microglomerulus (arrow). **e**, Electroporated dye labels a single PN ($n = 684$), which has a bouton that innervates the targeted microglomerulus (arrow). Note that the other KCs that synapse on this PN bouton were not labelled in this example. **f**, The photolabelled claw ensheathes the red dye-labelled PN bouton. Scale bar, 5 μm. **g**, The photolabelled KC projects to the α/β lobes of the MB whereas the dye-labelled PN innervates the DM6 glomerulus. **h**, A photolabelled KC with three claws. **i**, Three PNs innervating the DA1, VC4 and DL3 glomeruli are labelled upon loading all the claws of the KC depicted in **h**. Soma of the DA1 PN and VC4 PN are outlined whereas the DL3 soma is out of the plane. All scale bars are 10 μm except where noted.

neurons in the antennae[19]. The remaining 18 PNs innervated different uncharacterized regions of the brain.

We observe that the distribution of the glomerular inputs to KCs is not uniform (Fig. 3). Inputs from the DA1 and DC3 glomeruli are most frequent, with each accounting for 5.1% of the total connections (Fig. 3). The non-uniform distribution reflects the fact that the size and number of calycal boutons formed by PNs varies across glomeruli (Supplementary Fig. 1). For instance, the PNs associated with the DA1 and DC3 glomeruli form more numerous boutons in comparison with the PNs of less frequently represented glomeruli (Supplementary Fig. 1). We also observe that there is a small but significant difference between the inputs to the α/β and γ KCs ($P < 0.001$) (Supplementary Figs 2 and 3). All subsequent statistical analyses were therefore performed separately on both the α/β and γ data sets, but failed to reveal any significant difference between the two data sets. Therefore, only the results obtained from the full data set are shown.

Statistical analyses of the 665 connections allow us to search for structure among the connections between glomeruli and KCs. First, we determined whether the KCs receiving input from a given glomerulus have a higher probability of receiving additional input from that same glomerulus. Of the 200 KCs in the data set, only 11 receive two inputs from the same glomerulus, and none receive three or more such inputs (Figs 3, 4a, b and Supplementary Table 1). We determined whether the frequency of convergent input from a single glomerulus is significantly above or below chance expectations by randomly shuffling the connections in the data set between the different KCs, while maintaining the number of connections each of them receives. This shuffling maintains the frequency of glomerular connections observed in the experimental data, but eliminates any potential, non-random patterns of inputs onto individual KCs. This shuffling is used in all subsequent statistical analyses. The frequency of multiple connections from the same glomerulus in the observed and shuffled data sets is not significantly different (Fig. 4b). Thus, we observe no KCs that receive preferential inputs from a single glomerulus. Rather, individual KCs integrate information from multiple different glomeruli.

We next determined whether KCs are connected to any preferential pair, trio or quartet of glomeruli. Of the 1,378 ($53 \times 52/2$) different pairs of glomeruli that could converge onto an individual KC, 508 distinct pairs appear in the data set (Fig. 4a). 310 of these pairs connect to only one of the 200 KCs analysed, whereas certain pairs of glomeruli connect to multiple KCs (Fig. 4a, c). The DA1–DC3 pair, for example, converges onto nine different KCs (Fig. 3 and Supplementary Table 1). There are combinations of glomerular trios that connect with two different KCs, and one case in which two KCs receive inputs from

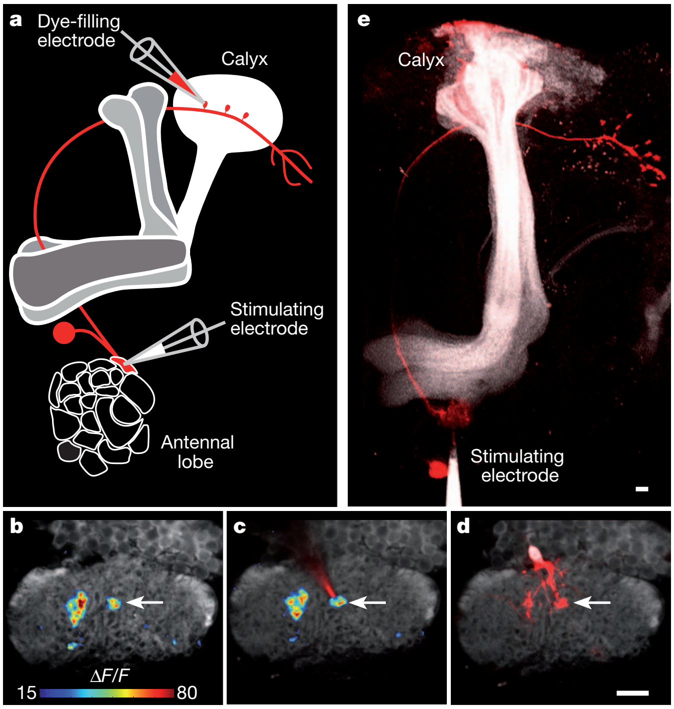


Figure 2 | Dye labelling identifies functional connections between PNs and KCs. **a**, Schematic illustration of the strategy used to identify functional connections between PNs and KCs. An AL glomerulus (here DL3) is stimulated by local iontophoresis of acetylcholine (stimulating electrode). Optical recordings of calcium-mediated changes in fluorescence ($\Delta F/F$) are measured in the MB calyx of a fly expressing GCaMP3 driven by the KC-specific promoter OK107^{GAL4}. A microglomerulus activated by the stimulation of DL3 is targeted for dye electroporation, identifying the pre-synaptic PN (red). **b**, Stimulation of the DL3 glomerulus activates several microglomeruli dispersed through the calyx. **c**, An electrode filled with Texas Red dextran is positioned into the centre of an activated microglomerulus (arrow) highlighted by the recorded $\Delta F/F$. **d**, Electroporation of dye into the targeted microglomerulus labels a single PN bouton (arrow). **e**, The labelled bouton extends from a single dye-filled PN that innervates the stimulated DL3 glomerulus ($n = 10$). Note that the stimulating electrode is visualized by addition of Alexa-488 dextran dye to the acetylcholine. Scale bars are 10 µm.

the same quartet of glomeruli (Fig. 3 and Supplementary Table 1). However, the observed frequency with which the different pairs, trios and quartets converge onto different KCs is consistent with expectations from the shuffled data set (Fig. 4c and data not shown). Thus, the identity of a glomerulus connected to a KC provides no predictive information as to the identity of the remaining glomerular inputs onto that neuron.

Glomeruli can be grouped based on biological properties shared by their associated OSNs (sensilla type, odour specificity) and PNs (developmental origin and topography of their axonal projections). KCs might receive preferential input from one or another of these glomerular categories. For example, the OSNs innervating the AL are derived from three sensillar types (basiconic, coeloconic and trichoid sensilla) that project to three classes of glomeruli tuned to different odour categories[20]. If individual KCs were tuned to a particular class of odours, they might preferentially integrate inputs from one type of sensillum. Statistical analyses, however, reveal that KCs that receive an input from one sensillar type are no more or less likely to receive additional inputs from this or any other type of sensillum than is predicted by chance (Fig. 4e–h). Sensillar type, however, provides only a coarse correlate of odour tuning. Therefore, we also grouped glomeruli based on the similarity of their odour response profiles[21] and again observed no structure in the inputs to a KC that correlated with odour tuning (Fig. 4d and Supplementary Fig. 4).

We also classified glomeruli on the basis of the properties of their PNs. PN axons from different glomeruli project to broad but stereotyped domains in the lateral horn and calyx of the MB[9]. Input to an individual KC could be shaped by the topography of PN projections. Analysis of the distribution of inputs to a given KC, however, fails to reveal any preferential PN connectivity that reflects the organization of their projection in either the MB calyx (Fig. 4d and Supplementary Fig. 5) or lateral horn (Fig. 4d and Supplementary Fig. 6). KCs do not preferentially integrate information from glomeruli innervated by PNs sharing a developmental origin[22] (Fig. 4d and Supplementary Fig. 7). In addition, KCs do not select their input on the basis of topographical constraints as suggested by a previous study[23] (Fig. 4d and Supplementary Fig. 8). Finally, three glomeruli are innervated by Fru-expressing OSNs and PNs[16,17]. We do not observe preferential pairing of inputs from Fru^{+} PNs onto individual KCs (Fig. 4d and Supplementary Fig. 9). Moreover, although most γ KCs express Fru, there is no preferential input from Fru^{+} glomeruli to γ KCs (Supplementary Fig. 2).

Next, we performed an unbiased search for structure by examining correlations within the connectivity matrix between the 53 glomeruli (51 AL glomeruli and 2 pseudoglomeruli) and the 200 KCs. Correlations were extracted by performing a principal component analysis of this matrix (Supplementary Figs 10, 11). This analysis failed to

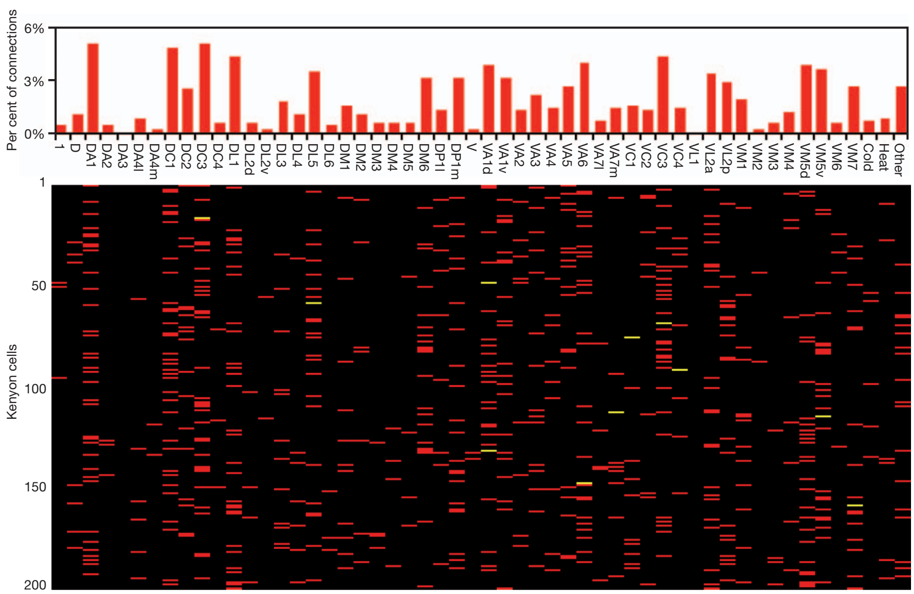


Figure 3 | The connectivity matrix between AL glomeruli and KCs. The 665 connections between the AL glomeruli and KCs are represented in a matrix in the lower panel. Each row corresponds to one of the 200 photolabelled KCs whereas each column refers to the 51 AL glomeruli, the two thermosensing pseudoglomeruli and the other uncharacterized brain regions. Glomeruli connected once to a given KC are depicted as red bars. Glomeruli connected twice to the same KC are labelled as yellow bars. In the upper panel, the connections to all glomeruli and other brain regions are sorted according to their observed frequency. The upper panel is a histogram of the frequency of occurrence for each input source.

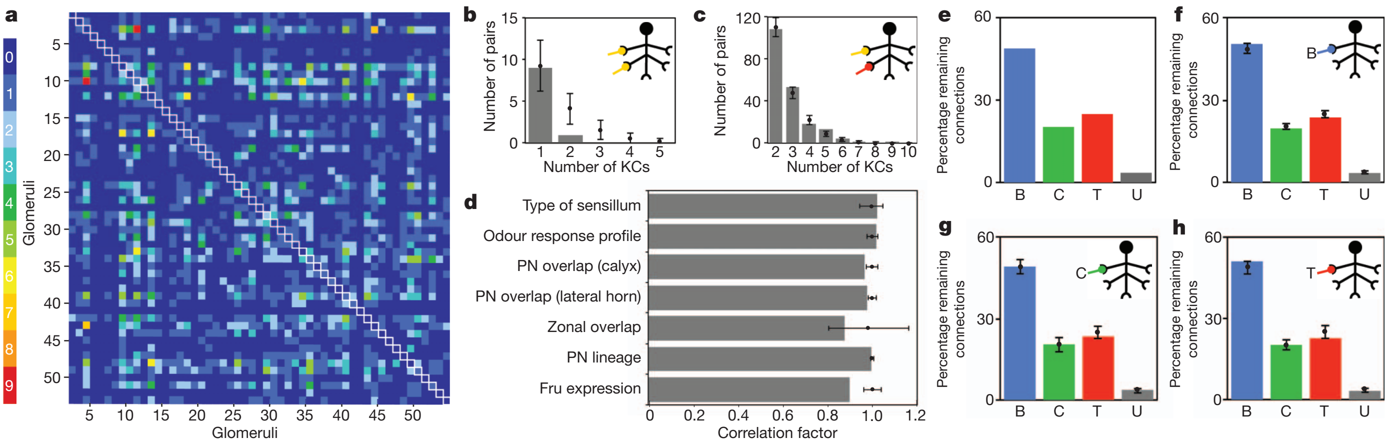


Figure 4 | KCs do not receive structured input. **a**, Two glomeruli projecting to the same KC are considered a connected pair. All possible connected pairs are depicted as squares in a 53 × 53 matrix (51 AL glomeruli and 2 pseudoglomeruli), coloured according to their observed frequency in the data (white outlined squares along the diagonal depict the frequency of identical pairs where a glomerulus is paired with itself). **b**, The frequency of KCs receiving two connections from the same glomerulus (an identical pair, grey bars) is compared to the frequency of such cells in 1,000 shuffled data sets (error bars, ± s.d.). **c**, The frequency of KCs receiving input from the same non-identical pair (grey bars) is compared to the frequency of such cells in 1,000 shuffled data sets (error bars, ± s.d.). **d**, Glomeruli are grouped based upon different anatomic or functional parameters[7,16,17,20–23]. For each listed parameter, the percentage of connections across KCs receiving at least one input from a given group (as shown in **f**, **g** and **h** for type of sensilla) is divided by the corresponding percentage observed in the full data set (as shown in **e**). A value of 1 for this quotient would indicate that the distributions across the selected KC groups and the full data set are identical. All analyses were also performed on 1,000 shuffled data sets (black circles, ± s.d.). **e**, The glomerular connections in the data set are grouped according to whether they receive input from an OSN that innervates a basiconic (blue), coeloconic (green), trichoid (red) or uncharacterized sensillum (grey). **f**–**h**, The distribution of the remaining glomerular connections to the 168 KCs receiving at least one input from a basiconic glomerulus (**f**), the 104 KCs receiving at least one input from a coeloconic glomerulus (**g**), and the 125 KCs receiving at least one input from a trichoid glomerulus (**h**) are shown. The frequency in 1,000 shuffled data sets are shown (black circles, average; error bars ± s.d.).

reveal structure in the input to KCs other than that inherent in the non-uniform distribution of glomerular inputs.

These data are consistent with a model in which each KC receives input from a combination of glomeruli randomly chosen from the non-uniform distribution of glomerular projections to the MB. Classification of either glomeruli or KCs on the basis of several shared developmental, anatomic and functional features fails to reveal structured input onto individual KCs. Members of a given PN class do not preferentially converge onto an individual KC, nor do members of a KC class receive specific and distinguishing PN inputs. A given KC can integrate information from glomeruli activated by food odours, pheromones, CO_2 and even temperature. Recent data indicates that the extrinsic output neurons of the MB that are responsible for the different forms of learned behaviour are anatomically segregated and synapse with KC axons within a specific MB lobe[13,14]. Interestingly, similar glomerular inputs are observed for the KCs that innervate the different lobes of the MB. This random input to individual KCs provides a mechanism to contextualize a rich diversity of novel KC responses.

It is important to note that the tracing procedure we have developed only allows us to characterize the inputs to a single KC per fly. It is therefore possible that the inputs to every KC are determined, but this developmental program results in a distribution of glomerular inputs that appears random. However, it is difficult to conceive of a development mechanism that could dictate the identity of inputs to each of the seven claws of the 2,000 KCs. Moreover, the logic of employing complex and unlikely identity codes to achieve an uncorrelated distribution of inputs is elusive. Indeed, a previous study examined the electrophysiological response of KCs to different odours in a line of flies that labels only 23 α/β neurons but failed to identify replicate KCs with shared odour response profiles[24]. These observations support the conclusion that the complement of glomerular inputs to KCs differs in different individuals. In addition, we cannot, from the analysis of the glomerular inputs to 200 KCs, exclude the existence of small subsets of KCs that received determined inputs from the AL. Nonetheless, our data are most consistent with a model in which the majority of individual KCs receive input from a random collection of glomeruli, a finding with important implications for odour processing in the MB.

If the connections from AL to MB are indeed random, a given odour will activate a different ensemble of KCs in different flies. However, in an individual fly, a given odour will consistently activate the same ensemble. This representation must acquire valence through experience or unsupervised activity-dependent plasticity to dictate an appropriate behavioural output. Uncorrelated glomerular input to KCs affords the fly with the ability to impart meaning to a diversity of novel and unpredictable sensory stimuli that it may encounter throughout its life. Plasticity at highly convergent synapses between KC axons and MB extrinsic neurons could mediate experience-dependent behavioural output, an elemental feature of MB function. Thus, the fly has evolved an olfactory circuit with a connectivity that optimizes its ability to contextualize and respond appropriately to a rich array of olfactory experiences.

METHODS SUMMARY

Photolabelling of individual KCs were performed on flies expressing UAS–C3PA–GFP and/or UAS–SPA–GFP under the control of either the synaptobrevin^{GAL4} or Fruitless^{GAL4} promoters. Texas-red dextran dye (Invitrogen) was electroporated into the centre of a photolabelled KC claw to label the PN connected to it. The details on this tracing technique and all other procedures are provided in the full methods section.

Full Methods and any associated references are available in the online version of the paper.

Received 28 September; accepted 11 March 2013.
Published online 24 April 2013.

1. Heisenberg, M. Mushroom body memoir: from maps to models. *Nature Rev. Neurosci.* **4,** 266–275 (2003).
2. Turner, G. C., Bazhenov, M. & Laurent, G. Olfactory representations by *Drosophila* mushroom body neurons. *J. Neurophysiol.* **99,** 734–746 (2008).
3. Honegger, K. S., Campbell, R. A. & Turner, G. C. Cellular-resolution population imaging reveals robust sparse coding in the *Drosophila* mushroom body. *J. Neurosci.* **31,** 11772–11785 (2011).
4. Wang, J. W., Wong, A. M., Flores, J., Vosshall, L. B. & Axel, R. Two-photon calcium imaging reveals an odor-evoked map of activity in the fly brain. *Cell* **112,** 271–282 (2003).
5. Ng, M. *et al.* Transmission of olfactory information between three populations of neurons in the antennal lobe of the fly. *Neuron* **36,** 463–474 (2002).

6. Wong, A. M., Wang, J. W. & Axel, R. Spatial representation of the glomerular map in the *Drosophila* protocerebrum. *Cell* **109,** 229–241 (2002).
7. Marin, E. C., Jefferis, G. S., Komiyama, T., Zhu, H. & Luo, L. Representation of the glomerular olfactory map in the *Drosophila* brain. *Cell* **109,** 243–255 (2002).
8. de Belle, J. S. & Heisenberg, M. Associative odor learning in *Drosophila* abolished by chemical ablation of mushroom bodies. *Science* **263,** 692–695 (1994).
9. Jefferis, G. S. *et al.* Comprehensive maps of *Drosophila* higher olfactory centers: spatially segregated fruit and pheromone representation. *Cell* **128,** 1187–1203 (2007).
10. Butcher, N. J., Friedrich, A. B., Lu, Z., Tanimoto, H. & Meinertzhagen, I. A. Different classes of input and output neurons reveal new features in microglomeruli of the adult *Drosophila* mushroom body calyx. *J. Comp. Neurol.* **520,** 2185–2201 (2012).
11. Leiss, F., Groh, C., Butcher, N. J., Meinertzhagen, I. A. & Tavosanis, G. Synaptic organization in the adult *Drosophila* mushroom body calyx. *J. Comp. Neurol.* **517,** 808–824 (2009).
12. Yasuyama, K., Meinertzhagen, I. A. & Schurmann, F. W. Synaptic organization of the mushroom body calyx in *Drosophila melanogaster. J. Comp. Neurol.* **445,** 211–226 (2002).
13. Tanaka, N. K., Tanimoto, H. & Ito, K. Neuronal assemblies of the *Drosophila* mushroom body. *J. Comp. Neurol.* **508,** 711–755 (2008).
14. Séjourné, J. *et al.* Mushroom body efferent neurons responsible for aversive olfactory memory retrieval in *Drosophila. Nature Neurosci.* **14,** 903–910 (2011).
15. Aso, Y. *et al.* The mushroom body of adult *Drosophila* characterized by GAL4 drivers. *J. Neurogenet.* **23,** 156–172 (2009).
16. Manoli, D. S. *et al.* Male-specific fruitless specifies the neural substrates of *Drosophila* courtship behaviour. *Nature* **436,** 395–400 (2005).
17. Stockinger, P., Kvitsiani, D., Rotkopf, S., Tirian, L. & Dickson, B. J. Neural circuitry that governs *Drosophila* male courtship behavior. *Cell* **121,** 795–807 (2005).
18. Keleman, K. *et al.* Dopamine neurons modulate pheromone responses in *Drosophila* courtship learning. *Nature* **489,** 145–149 (2012).
19. Gallio, M., Ofstad, T. A., Macpherson, L. J., Wang, J. W. & Zuker, C. S. The coding of temperature in the *Drosophila* brain. *Cell* **144,** 614–624 (2011).
20. Vosshall, L. B. & Stocker, R. F. Molecular architecture of smell and taste in *Drosophila. Annu. Rev. Neurosci.* **30,** 505–533 (2007).
21. Hallem, E. A. & Carlson, J. R. Coding of odors by a receptor repertoire. *Cell* **125,** 143–160 (2006).
22. Yu, H. H. *et al.* A complete developmental sequence of a *Drosophila* neuronal lineage as revealed by twin-spot MARCM. *PLoS Biol.* **8,** e1000461 (2010).
23. Lin, H. H., Lai, J. S., Chin, A. L., Chen, Y. C. & Chiang, A. S. A map of olfactory representation in the *Drosophila* mushroom body. *Cell* **128,** 1205–1217 (2007).
24. Murthy, M., Fiete, I. & Laurent, G. Testing odor response stereotypy in the *Drosophila* mushroom body. *Neuron* **59,** 1009–1023 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Bargmann, T. Jessell, F. Maderspacher, L. Paninski and members of the Axel laboratory for comments on the manuscript; C. Franqui for assistance with fly work; and P. Kisloff, M. Gutierrez and A. Nemes for assistance with general laboratory concerns and the preparation of this manuscript. This work was funded in part by a grant from the Foundation for the National Institutes of Health through the Grand Challenges in Global Health Initiative (R.A.). Further financial support was provided by the Howard Hughes Medical Institute (R.A.), by the Swartz and Gatsby Foundations (L.F.A.) and by the Pew Charitable Trusts, McKnight Foundation, and New York Stem Cell Foundation (V.R.).

Author Contributions S.J.C.C., V.R., L.F.A. and R.A. planned the research and wrote the paper; S.J.C.C. and V.R. performed the experiments; L.F.A. performed all statistical analyses.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.A. (ra27@columbia.edu).

METHODS

Fly stocks. All fly transgenic lines (synaptobrevin^{GAL4}, OK107^{GAL4}, GH146^{GAL4}, Fruitless^{GAL4}, UAS–C3PA–GFP, UAS–SPA–GFP and UAS–GCaMP3) have been described previously[25–28].

Imaging. All imaging experiments were performed using an Ultima two-photon laser scanning microscope (Prairie Technologies) equipped with galvanometers driving a Chameleon XR laser (Coherent). Emitted photons were collected with a GaAsP photodiode detector (Hamamatsu) or a PMT detector through a $\times$60 0.9 numerical aperture water immersion objective (Olympus). All images were acquired at a resolution of 512 by 512 pixels using 1 μm intervals between optical slices.

Photolabelling neurons through photoactivation of PA-GFP. Individual neurons were photolabelled by converting PA-GFP under the guidance of a two-photon microscope. In an initial set of experiments, individual KC soma were randomly selected for photolabelling in flies expressing one copy of UAS–SPA and two copies of UAS–C3PA[25] under the control of the pan-neuronal promoter synaptobrevin^{GAL4}. The number of γ KCs labelled using this procedure was lower than expected. This most likely reflects the fact that the somas of these neurons are buried at the core of the MB, a region less accessible to photoactivation. We therefore corrected for this bias by generating a second set of experiments using a γ KCs specific promoter. In this set, individual γ KCs were photolabelled in flies expressing two copies of UAS–C3PA under the control of the Fruitless^{GAL4}. In this set of experiments, flies also expressed C3PA–GFP in the majority of the PNs using the GH146^{GAL4} promoter to facilitate the identification of dye-labelled glomeruli in the AL. In all experiments, the brains of 1–3 days old male flies isolated from females after eclosion were dissected in saline (108 mM NaCl, 5 mM KCl, 2 mM $CaCl_2$, 8.2 mM $MgCl_2$, 4 mM $NaHCO_3$, 1 mM NaH_2PO_4, 5 mM trehalose, 10 mM sucrose, 5 mM HEPES pH 7.5, osmolarity adjusted to 275 mOsm) and incubated in 2 mg ml^{-1} collagenase (Sigma-Aldrich) for approximately 1 min. Brains were pinned on a thin sheet of Sylgard (World Precision Instruments) placed at the bottom of an imaging chamber filled with saline. The MB was first imaged at 925 nm (a wavelength at which photoconversion is relatively inefficient) in order to define a region of interest over the soma of the targeted KC. PA-GFP was subsequently photoactivated within that region through consecutive exposures to 710 nm laser light (a wavelength that efficiently photoconverts the fluorophore). A resting period of 10 min was allowed for the photoconverted fluorophore to diffuse into the distal KC processes. Photoactivation power was typically between 9 and 30 mW measured at the back of the objective and exposure time was less than 1 s. KCs that extend dendritic claws only to the accessory calyx but not the main olfactory calyx were found but these KCs were not used in this study. All the PNs innervating a particular glomerulus were photolabelled in flies expressing one copy of UAS–SPA and two copies of UAS–C3PA under the control of the pan-neuronal driver synaptobrevin^{GAL4} by targeting laser light at a region within a given glomerulus using a similar strategy (photoactivation power: between 25 and 32 mW, photoactivation exposure time: on average 1 min, resting time: 30 min).

Electroporation of dye into the PN connected to a KC claw. The PNs connected to a given photolabelled KC were identified by sequentially electroporating 100 mg ml^{-1} 3,000-Da Texas-red dextran dye (Invitrogen) into individual KC claws. Glass electrodes (Sutter Instruments) were pulled to a resistance of 9–11 MΩ. Each electrode was fire-polished using a micro-forge (Narishige) to narrow its opening. Electrodes were back-filled with the dye. Under the guidance of a two-photon microscope, the electrode was centred into a photolabelled KC claw using motorized manipulators (Sutter Instruments). Short current pulses (each 30–50 V for 0.5 ms) were used to electroporate the dye into the PN bouton connected to the targeted claw. Although small claws are as easy to fill as larger ones, filling all the claws formed by a given KC was technically challenging because the photolabel weakens as more electroporations are performed on the same neuron. In less than 5% of experiments, the number of labelled PNs was larger than the number of claws filled and these samples were discarded. A smaller number of samples had fewer labelled PNs than expected, most probably because one or more of the labelled claw(s) were not properly filled. A *z*-stack of the entire AL was taken at the end of the experiment. Dye-labelled glomeruli were identified using the basal fluorescence of PA-GFP expressed in all or most PNs (under the control of the pan-neuronal promoter synaptobrevin or the PN specific GH146 promoter). Dye-labelled glomeruli were identified on the basis of their stereotyped position and shape in the AL, as well as the location of the soma of their associated PNs and whether they were GH146^{+}. The glomeruli connected to each of the 200 KCs were analysed by the same person. First, the 654 labelled glomeruli were identified in the 200 AL. This identification was repeated without consideration to the previous designation and the mismatch rate between the two scores was determined. This procedure was repeated until the mismatch rate was lower than 5% (about 15 rounds). This approach permitted the observer to become extremely familiar with characteristic properties of the glomerular map in the AL so that the mismatch rate diminished considerably in later rounds. In addition, the same identified glomeruli were compared across samples to ascertain that a given glomerulus displays the same shape and general location in all ALs. Although we cannot exclude that some glomeruli might have been misidentified, such error will most probably be consistent across the data set and thus should not alter the conclusion of our study.

Functional imaging. Optical imaging experiments were performed in flies expressing two copies of UAS–GCaMP3 under the control of the KC specific OK107^{GAL4} promoter. The brains of 1–2 days old male flies were dissected, desheathed, and pinned on a thin sheet of Sylgard in an imaging chamber filled with saline. Single glomeruli were stimulated as previously described[25]. In summary, glass electrodes were pulled to a resistance of 7–8 MΩ and filled with 2 mM acetylcholine (Sigma-Aldrich). 3,000 Da Alexa-488 dextran was added at 0.5 mg ml^{-1} to the acetylcholine to allow for fluorescent visualization of the stimulating electrode. The stimulating electrode was positioned into the centre of a superficial glomerulus and short current pulses (each 0.5–2 V for 500 ms) generated by a stimulator (Grass Technologies) allowed for selective and synchronous stimulation of the PNs innervating the impaled glomerulus. Images of the MB calyx were acquired at 925 nm at a frequency of 2 Hz. Activated KC claws within an individual microglomeruli displayed increases in fluorescence and were targeted with a glass electrode filled with Texas Red dye. Electroporation of the dye into the activated microglomerulus was performed as described in the previous section.

Image processing. Maximum-intensity projections of *z*-stacks were generated in ImageJ (NIH). In some experiments (for example, Fig. 1g), out of plane fluorescence at the surface of the brain arising from auto-fluorescence of the glial sheath was masked.

Statistical analyses. All experimentally derived results were compared with those obtained from 1,000 shuffled data sets. We generated the shuffled data by making a list of the glomeruli that contributed to the 665 connections in the data. We then randomly permuted this list and drew from it sequentially to construct a new set of connections for 200 model KCs, drawing as many random connections for each model KC as it receives in the experimental data. This shuffling maintains the frequency of glomerular connections and the number of connections per KC observed in the experimental data, but eliminates any potential, non-random patterns of inputs onto individual KCs.

Distribution of the glomerular inputs: the deviation from uniformity of the distribution of the 665 experimentally derived connections between glomeruli and KCs was quantified for the full data set as well as for different KC subpopulations using a χ^2 measure.

Pairwise analysis: the frequency of a KC receiving two inputs from a given pair of glomeruli was measured for all 1,378 different possible pairs (51 AL glomeruli and 2 thermosensing pseudo glomeruli). Pairwise analysis was performed on both the experimentally derived and shuffled data sets.

Group analyses: the 51 AL glomeruli were grouped according to different criteria (the two thermosensing pseudo glomeruli were grouped as 'uncharacterized' for these analyses). For each criterion, the distribution of the 665 experimentally derived connections was determined across all groups. KCs with at least one connection to a particular group were selected and the distribution of their remaining inputs was compared to that of the full data set. We performed similar analyses on the shuffled data. The distributions obtained from the experimental and shuffled data sets were compared to search for any positive or negative correlations that were statistically significant. This analysis examines whether the probability of a KC receiving a connection from a glomerulus of type A, P(A), is equal to the conditional probability P(A|B), conditioned on its receiving one input from a glomerulus of type B. The following criteria were tested:

Type of sensillum: glomeruli were clustered into three groups based on their sensillar origin[20].

Odour response profile: glomeruli were clustered into three groups based on the similarity of the odour response profiles of their associated OSNs as measured in a previous study[21].

PN overlap (calyx): glomeruli were clustered into four groups based on the overlap of their associated PNs in the calyx of the MB as measured in a previous study[9].

PN overlap (lateral horn): glomeruli were clustered into five groups based on the overlap of their associated PNs in the lateral horn as measured in a previous study[9].

Zonal overlap: glomeruli were clustered into five groups as defined by a previous study that found correlations in the topography of PN boutons and KC dendrites in the MB calyx[23].

PN lineage: glomeruli were clustered into three groups based on the developmental origin of their associated PNs as determined in a previous study[22].

Fru expression: glomeruli were clustered into two groups given that their associated OSNs and PNs both express (or not) Fru[16,17].

Singular value decomposition: singular value decomposition was performed on the experimentally derived connectivity matrix as well as on the shuffled data to

test for statistical significance. Any matrix can be expressed as the product of an orthogonal matrix, a diagonal matrix, and another orthogonal matrix in what is called a singular value decomposition. Large singular values, the non-zero elements of the diagonal matrix, correspond to structure detected in the original matrix. The corresponding rows or columns of the orthogonal matrices provide projections reflecting this structure. Examples of the sensitivity of this method for revealing structure in connectivity matrices are shown in Supplementary Fig. 11. The per cent variances reported are the squares of the singular values.

25. Ruta, V. *et al.* A dimorphic pheromone circuit in *Drosophila* from sensory input to descending output. *Nature* **468,** 686–690 (2010).
26. Pauli, A. *et al.* Cell-type-specific TEV protease cleavage reveals cohesin functions in *Drosophila* neurons. *Dev. Cell* **14,** 239–251 (2008).
27. Connolly, J. B. *et al.* Associative learning disrupted by impaired G_s signaling in *Drosophila* mushroom bodies. *Science* **274,** 2104–2107 (1996).
28. Stocker, R. F., Heimbeck, G., Gendre, N. & de Belle, J. S. Neuroblast ablation in *Drosophila* P[GAL4] lines reveals origins of olfactory interneurons. *J. Neurobiol.* **32,** 443–456 (1997).

LETTER

doi:10.1038/nature12057

Tension sensing by Aurora B kinase is independent of survivin-based centromere localization

Christopher S. Campbell[1] & Arshad Desai[1]

Accurate segregation of the replicated genome requires chromosome biorientation on the spindle. Biorientation is ensured by Aurora B kinase (Ipl1), a member of the four-subunit chromosomal passenger complex (CPC)[1,2]. Localization of the CPC to the inner centromere is central to the current model for how tension ensures chromosome biorientation: kinetochore–spindle attachments that are not under tension remain close to the inner centromere and are destabilized by Aurora B phosphorylation, whereas kinetochores under tension are pulled away from the influence of Aurora B, stabilizing their microtubule attachments[3–5]. Here we show that an engineered truncation of the Sli15 (known as INCENP in humans) subunit of budding yeast CPC that eliminates association with the inner centromere nevertheless supports proper chromosome segregation during both mitosis and meiosis. Truncated Sli15 suppresses the deletion phenotypes of the inner-centromere-targeting proteins survivin (Bir1), borealin (Nbl1), Bub1 and Sgo1 (ref. 6). Unlike wild-type Sli15, truncated Sli15 localizes to pre-anaphase spindle microtubules. Premature targeting of full-length Sli15 to microtubules by preventing Cdk1 (also known as Cdc28) phosphorylation also suppresses the inviability of Bir1 deletion. These results suggest that activation of Aurora B kinase by clustering either on chromatin or on microtubules is sufficient for chromosome biorientation.

All known mechanisms targeting the CPC to centromeric chromatin, in budding yeast and elsewhere, rely on the survivin (Bir1) subunit. Budding yeast CPC (Fig. 1a) is targeted by two Bir1-dependent mechanisms: interaction of Bir1 with Sgo1, which recognizes histone H2a phosphorylated by the kinetochore-localized kinase Bub1 (ref. 6), and direct binding of Bir1 to the Ndc10 (also known as Cbf2) subunit of the centromeric DNA-binding Cbf3 complex[7,8]. In other species, survivin binding to histone H3 phosphorylated on Thr 3 by haspin kinase is also implicated in CPC centromere targeting[9–11]; however, deletion of the two haspin-like genes (*ALK1* and *ALK2*) does not lead to a growth phenotype (see below), suggesting that this mechanism may not operate in budding yeast.

In agreement with the view that Bir1-directed targeting of the CPC to centromeres is critical for chromosome biorientation, the majority of *bir1*Δ spores fail to grow (Fig. 1b) and temperature-sensitive mutations in *bir1* show chromosome missegregation similar to that observed in *ipl1* and *sli15* mutants[12,13]. A low frequency (10%, $n = 60$) of *bir1*Δ spore survival is observed after extended incubation[14] (Fig. 1b); these survivors, which we refer to as *bir1*Δ*, have high chromosome missegregation rates and harbour severe aneuploidy (Supplementary Fig. 1a and data not shown). To determine whether the severe *BIR1* deletion phenotype is due to inability of the CPC to target to the inner centromere, we generated truncations of the Sli15 amino terminus that are predicted to eliminate the interaction of Sli15–Ipl1 with Bir1–Nbl1 (ref. 15). Surprisingly, truncations of up to 228 N-terminal amino acids of Sli15 (the longest viable truncation; referred to hereafter as Sli15(ΔNT); Fig. 1c) showed no lethality—cells harbouring these truncations as the sole source of Sli15 grew indistinguishably from wild type (Figs 1c and 2b). Further truncations that encroached on the Sli15 central domain were lethal (Fig. 1c). Immunoprecipitation experiments indicated that deleting the Sli15 N terminus eliminated the interaction with Bir1 (Fig. 1d). Analysis of CPC anaphase spindle localization, which is dependent on Sli15, confirmed this result. Whereas Sli15(ΔNT) and Ipl1 localized on the anaphase spindle (Fig. 1e), Bir1 and Nbl1 were delocalized in *sli15(ΔNT)* mutant cells (Fig. 1f).

The above results show that Sli15(ΔNT) is viable even though it disrupts CPC formation and disconnects the kinase activity of the CPC from the Bir1 subunit that targets it to centromeres. Consistent with this finding, *sli15(ΔNT)* fully suppressed inviability of *bir1*Δ and *nbl1*Δ: *sli15(ΔNT) bir1*Δ or *sli15(ΔNT) nbl1*Δ double-mutant spores formed colonies at the expected frequency with normal growth properties (Fig. 2a, b). Wild-type Sli15 and Sli15(ΔNT) were expressed at

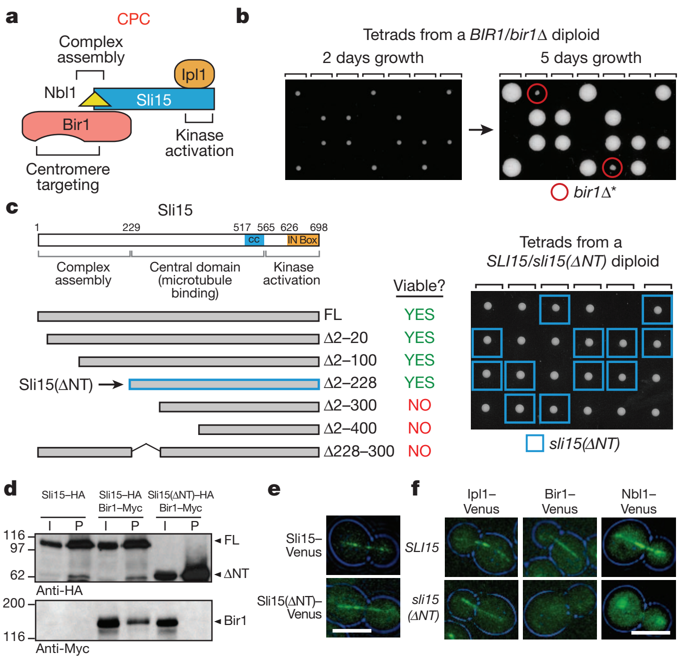


Figure 1 | Deletion of the Sli15 N terminus prevents association with Bir1 but does not affect cell viability or growth. **a**, Schematic of the CPC in budding yeast. Bir1, survivin; Ipl1, Aurora B kinase; Nbl1, borealin/dasra; Sli15, INCENP. **b**, Phenotype of *bir1*Δ. Tetrad dissections from a *bir1*Δ heterozygote. The four spores from individual tetrads are arrayed in columns. Rare survivors (*bir1*Δ*) are observed after extended growth (right). **c**, Phenotype of Sli15 truncations (left) and tetrad dissections from a *sli15(ΔNT)* heterozygote (right). **d**, Co-immunoprecipitation analysis of full-length (FL) Sli15 or Sli15(ΔNT) and Bir1. 9-Myc and 6-haemagglutinin (HA) tags were inserted at endogenous loci to generate functional C-terminally tagged proteins. Bir1 co-immunoprecipitates with full-length Sli15 but not with Sli15(ΔNT). I, input; P, anti-HA immunoprecipitate. **e**, Localization of Sli15–Venus and Sli15(ΔNT)–Venus to the anaphase spindle. Scale bar, 5 μm. **f**, Localization of CPC components during anaphase in cells with either wild-type Sli15 or Sli15(ΔNT). The cell outline is in blue. Scale bar, 5 μm.

[1]Ludwig Institute for Cancer Research and Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, California 92037, USA.

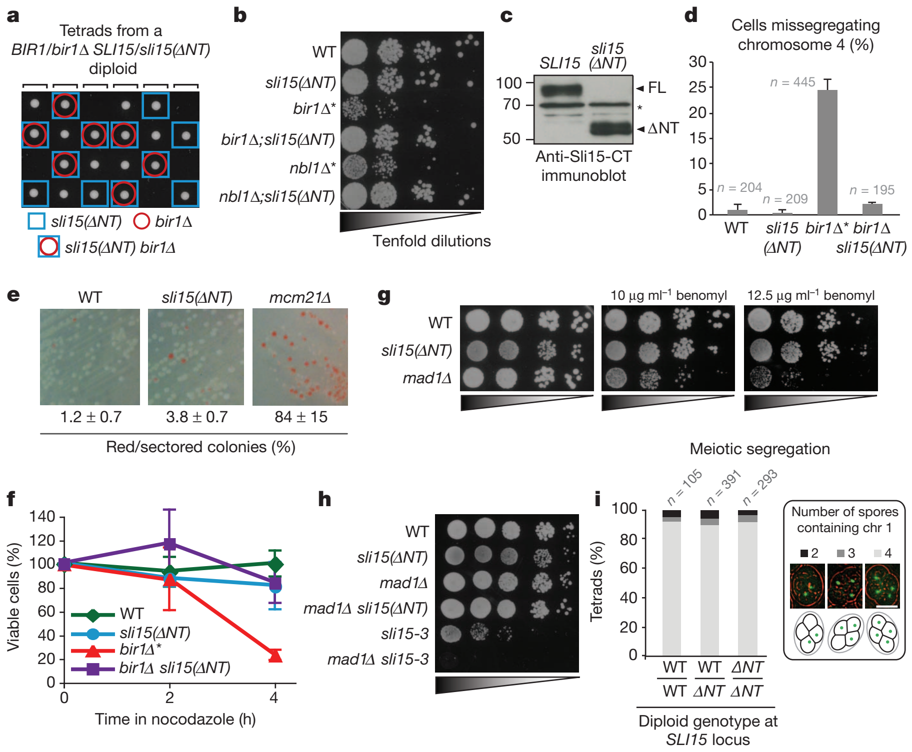


Figure 2 | *sli15(ΔNT)* suppresses lethality of *bir1Δ* and *nbl1Δ*, and shows normal fidelity mitotic and meiotic chromosome segregation. **a**, Tetrad dissection showing viability of *sli15(ΔNT) bir1Δ* double-mutant spores. **b**, Tenfold serial dilutions of cells of indicated genotypes. *bir1Δ** and *nbl1Δ** represent rare survivors recovered as shown in Fig. 1b. WT, wild type. **c**, Immunoblot of extracts prepared from strains expressing either wild-type Sli15 or Sli15(ΔNT) and blotted using an antibody raised against the C terminus (CT) of Sli15. Asterisk indicates a non-specific band recognized by the primary antibody that serves as a loading control. **d**, Analysis of segregation fidelity of green fluorescent protein-tagged chromosome 4 in cells of the indicated genotypes. The average of 3–5 experiments is shown; error bars represent standard error. **e**, Minichromosome loss assay. The percentage and standard error of colonies that were either red or sectored is shown. **f**, Viability after transient nocodazole treatment of cells of the indicated genotypes. The average of 2–4 experiments is shown; error bars represent standard error. **g**, Serial dilutions of cells with the indicated genotypes spotted on plates with different concentrations of the microtubule-depolymerizing drug benomyl. **h**, Lack of synthetic lethality/sickness following checkpoint inhibition in *sli15(ΔNT)* cells. Plates were incubated at 37 °C. *sli15-3* is a temperature-sensitive mutant that compromises Ipl1 activation[17]. **i**, Meiotic segregation after sporulation of diploid cells. The presence or absence of chromosome 1 was scored for each individual spore in a tetrad. Scale bar, 5 μm.

similar levels, indicating that suppression of *bir1Δ* and *nbl1Δ* lethality was not due to overexpression of mutant Sli15 (Fig. 2c). Sli15 was not detected in *bir1Δ** and *nbl1Δ** strains, suggesting that full-length Sli15 is destabilized when complex formation is disrupted (Supplementary Fig. 1b and Supplementary Discussion). We conclude that Sli15(ΔNT)–Ipl1 is sufficient to perform the essential function(s) of the CPC in the complete absence of Bir1 or Nbl1.

To assess the degree to which Sli15(ΔNT)–Ipl1 substitutes for the full CPC, we analysed chromosome segregation and biorientation using multiple assays. First, we monitored a single tagged chromosome, which revealed high-fidelity segregation for both the single *sli15(ΔNT)* and double *sli15(ΔNT) bir1Δ* mutants (Fig. 2d); by contrast, extensive missegregation was observed in the rare *bir1Δ** survivors (Fig. 2d and Supplementary Fig. 1a) and has been previously reported for *ipl1* kinase-activity-defective mutants[3,16,17]. Second, we monitored the segregation of a minichromosome (non-essential chromosome 3 fragment) in a sensitive colony colour assay, which revealed near-normal fidelity of segregation in *sli15(ΔNT)* cells (Fig. 2e; *mcm21Δ* cells, which are also viable and do not show significant growth defects but reduce chromosome segregation fidelity, are shown for comparison). Third, modest defects in chromosome biorientation are enhanced by transient nocodazole treatment and removal, due to the formation of multiple incorrect attachments during the recovery of collapsed spindles[18]. We therefore tested cell viability after transient nocodazole treatment. *bir1Δ** spores showed rapid death following transient nocodazole treatment; by contrast, *sli15(ΔNT)* and *sli15(ΔNT) bir1Δ* cells behaved similarly to wild-type cells (Fig. 2f). Fourth, mutations that increase chromosome segregation errors show sensitivity to microtubule-depolymerizing drugs such as benomyl[18,19]. *sli15(ΔNT)* growth rates were similar to wild type in the presence of benomyl (Fig. 2g). Fifth, mild defects in CPC activity, as well as deletion mutants of non-essential chromosome segregation proteins, show strong synthetic lethal/slow growth interactions with mutations in the spindle checkpoint (such as *mad1Δ* or *mad2Δ*) (for example, see ref. 20). *sli15(ΔNT)* mutant cells did not exhibit a synthetic lethal/slow growth interaction with *mad1Δ* (Fig. 2h). Sixth, *sli15(ΔNT)* mutant cells delayed cell cycle progression in response to lack of sister chromatid cohesion, indicating that they are competent to communicate absence of tension to the spindle checkpoint (Supplementary Fig. 2a). Last, subtle defects in chromosome segregation are often magnified during meiosis[21], but *sli15(ΔNT)* cells did not exhibit increased meiotic missegregation relative to control cells (Fig. 2i). Thus, *sli15(ΔNT)* cells show remarkably normal chromosome segregation fidelity during mitosis and meiosis in budding yeast.

We next tested whether Sli15(ΔNT) also bypasses mutations in the Bir1 localization pathway. Bir1-dependent targeting of the CPC to the inner centromere involves recognition of phosphorylated histone H2a by Sgo1 (Fig. 3a). Chromosome biorientation errors in *bub1Δ* and *sgo1Δ* mutants lead to severe growth defects, sensitivity to benomyl and difficulty correcting defective attachments after nocodazole washout[18]. *sli15(ΔNT)* suppressed the severe growth defect of both *bub1Δ* and *sgo1Δ* cells (Fig. 3b); in addition, the benomyl sensitivity, the rapid loss of viability after transient nocodazole treatment and the chromosome missegregation of *sgo1Δ* cells were also suppressed (Fig. 3b, c and Supplementary Fig. 1c). The benomyl sensitivity of *bub1Δ* was not suppressed as the spindle checkpoint function of Bub1 is independent of its role in CPC localization. Although haspin kinases contribute to CPC targeting in other organisms through the creation of a binding site on centromeric chromatin for survivin, deletion of the two haspin homologues in budding yeast had no growth phenotype on their own or in combination with *sli15(ΔNT)* (Fig. 3d).

An *ipl1* temperature-sensitive mutant is synthetic lethal with deletion mutants of the Ctf19 and Mcm21 subunits of the Ctf19 kinetochore complex, which provides a non-essential function in centromeric cohesion[20]. In contrast to the *bir1Δ*, *nbl1Δ*, *sgo1Δ* and *bub1Δ* mutants, whose lethality/severe growth defects were suppressed by *sli15(ΔNT)*, *ctf19Δ* and *mcm21Δ* mutants showed synthetic lethality/sickness with *sli15(ΔNT)* (Fig. 3e, f and Supplementary Fig. 2b, c). Combining

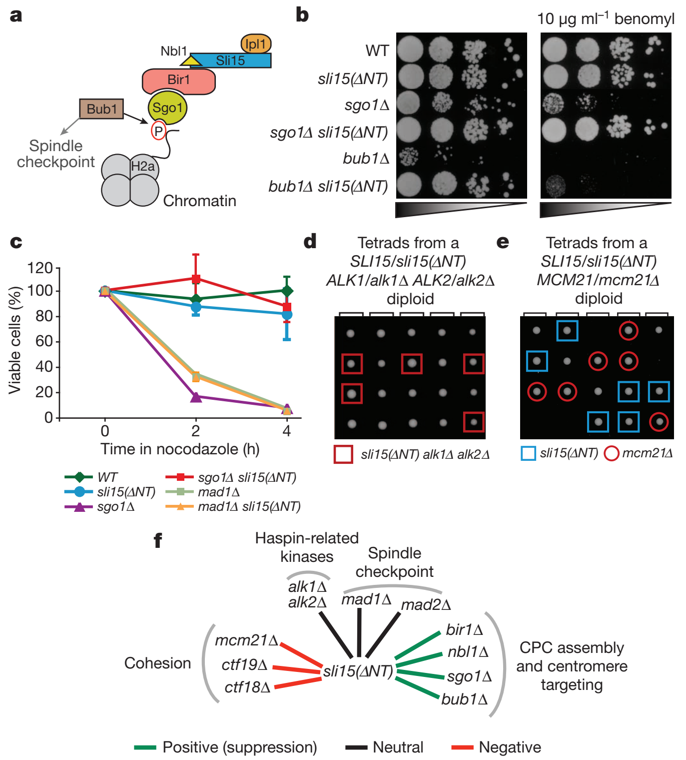


Figure 3 | *sli15(ΔNT)* suppresses mutants in the Bir1-dependent CPC-targeting pathway but is synthetically lethal with genes implicated in centromere cohesion. **a**, Schematic of the CPC-targeting mechanism involving Bub1 kinase and Sgo1. **b**, Suppression of *sgo1Δ* and *bub1Δ* growth phenotypes by *sli15(ΔNT)*. Compromised growth of *sgo1Δ* on benomyl is also suppressed by *sli15(ΔNT)*. **c**, Viability after transient nocodazole treatment of cells of the indicated genotypes. The average of 2–4 experiments is shown; error bars represent standard error. Wild-type and *sli15(ΔNT)* measurements are the same as in Fig. 2f. **d**, Tetrad analysis showing no synthetic defect for cells that are triple mutants for *sli15(ΔNT)*, *alk1Δ* and *alk2Δ*. **e**, Tetrad analysis showing synthetic lethality of *sli15(ΔNT)* and *mcm21Δ*. Similar results were observed for *ctf19Δ* (Supplementary Fig. 2b). **f**, Summary of genetic interactions shown by *sli15(ΔNT)*. A positive genetic interaction (green) indicates suppression; negative genetic interaction (red) indicates synthetic lethality/sickness and a neutral interaction (black) indicates lack of a synthetic phenotype.

sli15(ΔNT) with a deletion of *CTF18*, a separate non-essential mutant affecting cohesion establishment, also led to a synthetic sick phenotype (Supplementary Fig. 2b, c). Thus, whereas Bir1-dependent CPC targeting is dispensable for chromosome biorientation and segregation, this targeted CPC pool shows a functional connection with cohesion (Fig. 3f and Supplementary Discussion).

We next examined the localization of CPC subunits in wild-type cells and in cells expressing Sli15(ΔNT). In cells arrested in metaphase by depletion of the anaphase activator Cdc20, where the chromosomes are already bioriented, localization of Sli15 or Ipl1 on chromatin was not detected (Fig. 4b, top row and Supplementary Fig. 3b). However, localization of Sli15 and Ipl1 between sister kinetochore clusters (analogous to the inner-centromere localization in other species) was observed following brief microtubule depolymerization in asynchronously growing cells (Fig. 4a and Supplementary Fig. 3a). In *sli15(ΔNT)* cells, localization between sister kinetochore clusters was lost for both Sli15(ΔNT) and Ipl1 (Fig. 4a and Supplementary Fig. 3a); instead weak localization was observed coincident with kinetochore clusters (Fig. 4a and Supplementary Discussion). Thus, the Sli15(ΔNT)–Ipl1 complex supports accurate chromosome segregation without enriching between sister kinetochores *in vivo*.

In wild-type cells, the CPC is prevented from localizing to the spindle by Cdk1 phosphorylation until anaphase onset, when it is recruited to spindle microtubules and functions in spindle elongation[22–24]. However, the Sli15(ΔNT)–Ipl1 complex showed robust accumulation on the spindle in cells held in metaphase by depletion of the anaphase activator Cdc20 (Fig. 4b and Supplementary Fig. 3b). The central domain of Sli15—which harbours microtubule-binding activity—is required for chromosome biorientation[14] (Fig. 1c), and microtubule binding by human or frog INCENP activates Aurora B kinase through local clustering[25–27]. We therefore proposed that the Sli15(ΔNT) mutant may rely on clustering mediated by microtubule binding to activate Ipl1, which in turn detects and corrects defective kinetochore–spindle attachments. To test this idea, we wanted to determine whether

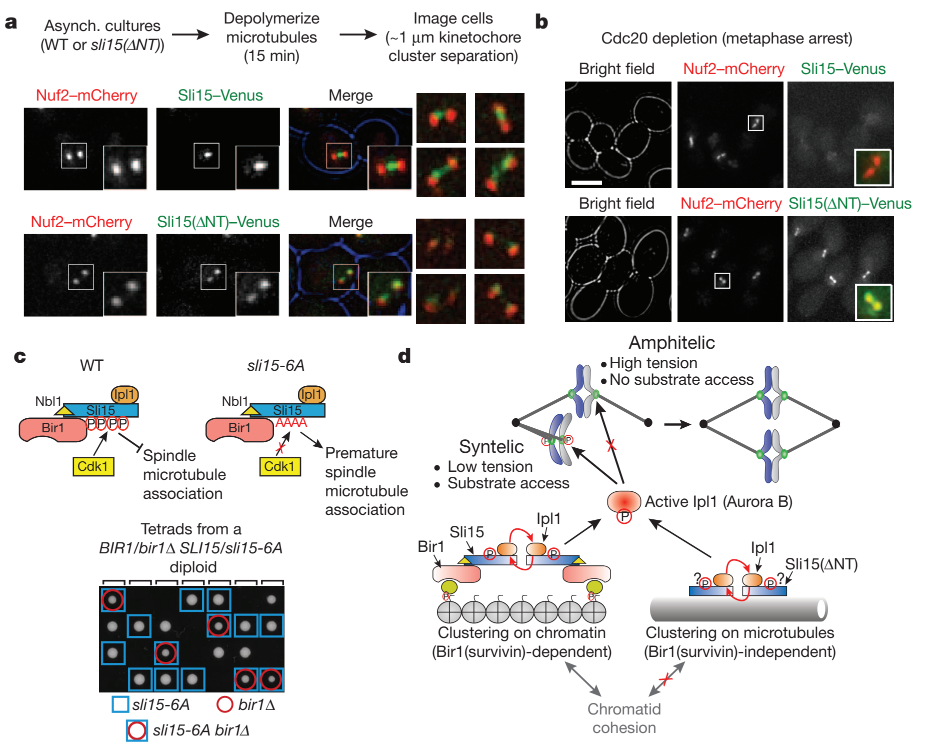


Figure 4 | Localization of Sli15(ΔNT) and relationship between Sli15 microtubule localization and suppression of *bir1Δ*. **a**, Images of cells expressing Nuf2–mCherry and either Sli15–Venus or Sli15(ΔNT)–Venus after brief microtubule depolymerization. A cell with ~1 μm separation of Nuf2–mCherry clusters is shown for each. Blue staining denotes cell outline. Boxes are 2.1 μm square. Four additional examples are shown on the right. See Supplementary Fig. 3a for similar analysis of Ipl1 localization. Asynch, asynchronous. **b**, Images of cells arrested in metaphase by Cdc20 depletion expressing Nuf2–mCherry and either Sli15–Venus or Sli15(ΔNT)–Venus. Scale bar, 5 μm. Merged insets are magnified 2.5-fold. See Supplementary Fig. 3b for similar analysis of Ipl1 localization. **c**, Schematic summarizing previous work on Cdk1 regulation of Sli15 spindle localization[22] and tetrad analysis showing growth of *sli15-6A bir1Δ* double-mutant cells. **d**, Model for mechanism of chromosome biorientation. Bir1-dependent chromatin clustering of the CPC or Bir1-independent clustering on microtubules of Sli15(ΔNT)–Ipl1 generates active Ipl1 kinase, which is capable of discriminating between correct and incorrect attachments.

prematurely clustering Sli15 on microtubules by another means bypassed the requirement for Bir1 for viability. For this purpose, we used the previously described *sli15-6A* mutant that prevents Cdk1 phosphorylation in the central domain of Sli15 and prematurely targets it to pre-anaphase spindle microtubules[22]. Consistent with our hypothesis, *sli15-6A* suppressed the inviability of *bir1Δ*: in contrast to *bir1Δ* alone (Fig. 1b) all double-mutant *sli15-6A bir1Δ* spores formed colonies (Fig. 4c). However, the *sli15-6A bir1Δ* double-mutant cells grew slowly and exhibited benomyl sensitivity, indicating that the suppression was partial, unlike the case for *sli15(ΔNT)* (not shown). As with *sli15(ΔNT)*, the suppression of *BIR1* deletion by *sli15-6A* was not due to overexpression of the mutant protein (Supplementary Fig. 3d). A Sli15 mutant that cannot be phosphorylated by Ipl1 and displays premature localization to the metaphase spindle[24] also partially rescued *BIR1* deletion (Supplementary Fig. 3c).

Our finding that biorientation occurs normally in the absence of Bir1-dependent targeting of the CPC has considerable implications for how the discrimination of correct (amphitelic; with tension) and incorrect (syntelic; lacking tension) attachments is achieved (Fig. 4d and Supplementary Discussion). Current models for biorientation emphasize the distance between inner-centromere-localized Aurora B kinase and outer-kinetochore-localized phosphatase activity as being critical for this discrimination[3–5]. Our findings instead suggest that active Aurora B, generated by clustering on either microtubules or centromeric chromatin, is capable of discriminating between correct and incorrect attachments and that this discrimination is intrinsic to the kinetochore (Fig. 4d). A parsimonious explanation for how this discrimination is achieved is substrate access, with correct, tense attachments becoming less sensitive to Aurora B activity. Super-resolution imaging studies have documented structural changes in microtubule-attached kinetochores under tension versus lacking tension[28], which may lead to changes in susceptibility to Aurora B phosphorylation. Defining the property that is detected by Aurora B to discriminate correct versus incorrect attachments should be facilitated by the finding that survivin-mediated centromere targeting of the CPC is not necessary for tension-sensing and chromosome biorientation in budding yeast.

METHODS SUMMARY

Yeast strains and media. All yeast strains used in this study are listed in Supplementary Table 1. Cells were cultured on yeast extract and peptone media with glucose unless otherwise indicated.

Microscopy. All microscopy was performed on a DeltaVision microscopy system (Applied Precision) and deconvolved with softWoRx software. Cells were prepared on 1% agar pads supplemented with complete synthetic media for imaging.

Nocodazole recovery assay. Cultures were synchronized with α-factor, washed and released into yeast extract peptone dextrose (YPD) with 10 μg ml^{-1} benomyl and 15 μg ml^{-1} nocodazole. Samples were collected every 2 h, and 100 μl of a 1:10,000 dilution was plated on YPD agar plates.

Immunoprecipitations. In total, 500 μl of cleared lysate was combined with 25 μl of antibody-bead slurry and rotated at 4 °C for 1 h. Beads were washed and resuspended in protein sample buffer for analysis by western blot.

Tetrad dissection. Spores were washed with water and digested with 1 mg ml^{-1} zymolyase for 5 min at 30 °C and then dissected onto YPD plates. Genotyping was performed by replica plating colonies onto selective media.

Full Methods and any associated references are available in the online version of the paper.

Received 6 November 2012; accepted 5 March 2013.
Published online 21 April 2013.

1. Ruchaud, S., Carmena, M. & Earnshaw, W. C. Chromosomal passengers: conducting cell division. *Nature Rev. Mol. Cell Biol.* **8,** 798–812 (2007).
2. Pinsky, B. A. & Biggins, S. The spindle checkpoint: tension versus attachment. *Trends Cell Biol.* **15,** 486–493 (2005).
3. Tanaka, T. U. *et al.* Evidence that the Ipl1-Sli15 (Aurora kinase-INCENP) complex promotes chromosome bi-orientation by altering kinetochore-spindle pole connections. *Cell* **108,** 317–329 (2002).
4. Liu, D., Vader, G., Vromans, M. J. M., Lampson, M. A. & Lens, S. M. A. Sensing chromosome bi-orientation by spatial separation of Aurora B kinase from kinetochore substrates. *Science* **323,** 1350–1353 (2009).
5. Lampson, M. A. & Cheeseman, I. M. Sensing centromere tension: Aurora B and the regulation of kinetochore function. *Trends Cell Biol.* **21,** 133–140 (2011).
6. Kawashima, S. A., Yamagishi, Y., Honda, T., Ishiguro, K.-I. & Watanabe, Y. Phosphorylation of H2A by Bub1 prevents chromosomal instability through localizing shugoshin. *Science* **327,** 172–177 (2010).
7. Yoon, H. J. & Carbon, J. Participation of Bir1p, a member of the inhibitor of apoptosis family, in yeast chromosome segregation events. *Proc. Natl Acad. Sci. USA* **96,** 13208–13213 (1999).
8. Cho, U.-S. & Harrison, S. C. Ndc10 is a platform for inner kinetochore assembly in budding yeast. *Nature Struct. Mol. Biol.* **19,** 48–55 (2012).
9. Yamagishi, Y., Honda, T., Tanno, Y. & Watanabe, Y. Two histone marks establish the inner centromere and chromosome bi-orientation. *Science* **330,** 239–243 (2010).
10. Kelly, A. E. *et al.* Survivin reads phosphorylated histone H3 threonine 3 to activate the mitotic kinase Aurora B. *Science* **330,** 235–239 (2010).
11. Wang, F. *et al.* Histone H3 Thr-3 phosphorylation by Haspin positions Aurora B at centromeres in mitosis. *Science* **330,** 231–235 (2010).
12. Shimogawa, M. M., Widlund, P. O., Riffle, M., Ess, M. & Davis, T. N. Bir1 is required for the tension checkpoint. *Mol. Biol. Cell* **20,** 915–923 (2009).
13. Makrantoni, V. & Stark, M. Efficient chromosome bi-orientation and the tension checkpoint in *Saccharomyces cerevisiae* both require Bir1. *Mol. Cell. Biol.* **29,** 4552–4562 (2009).
14. Sandall, S. *et al.* A Bir1-Sli15 complex connects centromeres to microtubules and is required to sense kinetochore tension. *Cell* **127,** 1179–1191 (2006).
15. Jeyaprakash, A. A. *et al.* Structure of a Survivin-Borealin-INCENP core complex reveals how chromosomal passengers travel together. *Cell* **131,** 271–285 (2007).
16. Biggins, S. & Murray, A. W. The budding yeast protein kinase Ipl1/Aurora allows the absence of tension to activate the spindle checkpoint. *Genes Dev.* **15,** 3118–3129 (2001).
17. Kim, J. H., Kang, J. S. & Chan, C. S. Sli15 associates with the ipl1 protein kinase to promote proper chromosome segregation in *Saccharomyces cerevisiae. J. Cell Biol.* **145,** 1381–1394 (1999).
18. Indjeian, V. B. The centromeric protein Sgo1 is required to sense lack of tension on mitotic chromosomes. *Science* **307,** 130–133 (2005).
19. Li, R. & Murray, A. W. Feedback control of mitosis in budding yeast. *Cell* **66,** 519–531 (1991).
20. Ng, T. M., Waples, W. G., Lavoie, B. D. & Biggins, S. Pericentromeric sister chromatid cohesion promotes kinetochore biorientation. *Mol. Biol. Cell* **20,** 3818–3827 (2009).
21. Shonn, M. A., McCarroll, R. & Murray, A. W. Requirement of the spindle checkpoint for proper chromosome segregation in budding yeast meiosis. *Science* **289,** 300–303 (2000).
22. Pereira, G. Separase regulates INCENP-Aurora B anaphase spindle function through Cdc14. *Science* **302,** 2120–2124 (2003).
23. Rozelle, D. K., Hansen, S. D. & Kaplan, K. B. Chromosome passenger complexes control anaphase duration and spindle elongation via a kinesin-5 brake. *J. Cell Biol.* **193,** 285–294 (2011).
24. Nakajima, Y. *et al.* Ipl1/Aurora-dependent phosphorylation of Sli15/INCENP regulates CPC–spindle interaction to ensure proper microtubule dynamics. *J. Cell Biol.* **194,** 137–153 (2011).
25. Sessa, F. *et al.* Mechanism of Aurora B activation by INCENP and inhibition by hesperadin. *Mol. Cell* **18,** 379–391 (2005).
26. Kelly, A. E. *et al.* Chromosomal enrichment and activation of the aurora B pathway are coupled to spatially regulate spindle assembly. *Dev. Cell* **12,** 31–43 (2007).
27. Tseng, B. S., Tan, L., Kapoor, T. M. & Funabiki, H. Dual detection of chromosomes and microtubules by the chromosomal passenger complex drives spindle assembly. *Dev. Cell* **18,** 903–912 (2010).
28. Wan, X. *et al.* Protein architecture of the human kinetochore microtubule attachment site. *Cell* **137,** 672–684 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors would like to thank the Desai and Oegema laboratories for discussions; S. Sandall, H. Hu and E. Manrriquez for assistance; B. Ren's laboratory for help with ChIP experiments; S. Biggins, D. Dawson, G. Pereira, G. Barnes, P. Hieter and the Yeast Resource Center for strains and plasmids; and K. Oegema, R. Green and J. DeLuca for comments on the manuscript. This work was supported by a National Institutes of Health (NIH) grant (GM074215) to A.D. and a Damon Runyon Cancer Research Foundation Fellowship (DRG 2007-09) to C.S.C. A.D. receives salary and other support from the Ludwig Institute for Cancer Research.

Author Contributions C.S.C. and A.D. designed experiments and wrote the manuscript. C.S.C. performed the experiments and analysed the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.D. (abdesai@ucsd.edu).

METHODS

Yeast strains and media. All yeast strains and plasmids used in this study are listed in Supplementary Table 1. Strains were grown in either yeast extract/peptone or synthetic media at 30 °C unless otherwise indicated. Epitope and fluorescent tags were inserted at the C terminus of genes at their native loci as previously described[29]. Gene truncations were made with the QuikChange Mutagenesis Kit (Agilent Technologies). Truncations and point mutants were integrated either at their native loci (by digesting the plasmid with the unique NruI site in the Sli15 promoter), or the *URA3* locus (by digesting with the StuI site in the *URA3* gene). For native locus integration, the wild-type copies of the gene were then excised by growing overnight in yeast extract peptone dextrose (YPD) and selecting for growth on 5-fluoroorotic acid (5-FOA) plates. All integrations were then checked by PCR and sequencing.

Immunoprecipitation and western blotting. Cells in exponential growth were pelleted and re-suspended in 600 µl immunoprecipitation buffer with protease inhibitors (50 mM Tris, pH 7.6, 150 mM NaCl, 1% Triton X-100, 1 mM EDTA, 2 mM phenylmethylsulphonyl fluoride, 4 mM benzamidine, cOmplete EDTA-free protease inhibitor cocktail (Roche)) and vortexed for 45 min with 400 µl glass beads. Lysates were cleared at 18,000*g* for 10 min, transferred to a new tube and centrifuged again. A total of 25 µl of antibody-bead slurry (Anti-HA monoclonal clone 3F10, Roche) was combined with 500 µl of cleared lysate and rotated at 4 °C for 1 h. Beads were washed once with immunoprecipitation buffer and three times with Tris buffer saline, pH 7.4, and then re-suspended in protein sample buffer. Samples were analysed by 8 or 10% SDS–PAGE and immunoblotted with anti-HA clone 3F10 (Roche) and anti-Myc mouse monoclonal 4A6 (Millipore), followed by horseradish peroxidase-conjugated secondary antibodies.

Nocodazole recovery assay. Yeast cultures were diluted to an attenuance ($D_{600\,nm}$) of 0.1 in YPD and incubated at 30 °C for 1.5 h. Next, 10 µM α-factor was added for 2.5 h. Cells were washed five times with YPD, re-suspended in YPD plus 10 µg ml^{-1} benomyl and 15 µg ml^{-1} nocodazole, and incubated at 23 °C. Samples were collected every 2 h, and 100 µl of a 1:10,000 dilution was plated on YPD agar plates. The percentage of colonies formed was determined by dividing the number of colonies at the indicated time point to the number at time zero after 3-day growth on plates. A minimum of 200 colonies were counted for each mutant at time zero.

Analysis of cell cycle progression after cohesin depletion. Overnight cultures of *GAL-MCD1* strains were diluted into fresh YPG media and arrested in G1 phase with 1 µM α-factor. Cells were then washed five times and released into fresh medium. A further 1 µM α-factor was added again when the cells had small buds to prevent entry into the next cell cycle. Samples were taken at the indicated time points and lysed by vortexing for 2 min with glass beads in sample buffer. The samples were then analysed by SDS–PAGE and western blot.

Microscopy. Overnight cultures were diluted ~100-fold and grown for 5 h. Cells were pelleted, washed once and re-suspended in water, placed on 1% agar pads supplemented with complete synthetic media, covered with a coverslip and sealed around the edges with VALAP (a 1:1:1 mixture of vaseline, lanolin (Fisher) and paraffin (Fisher) by weight). Images were collected on a DeltaVision microscopy system (Applied Precision) using a 100×, 1.3 NA Olympus U-PlanApo objective. Fourteen *z*-sections were taken with 0.5-µm steps and deconvolved with softWoRx software. Further image analyses, including maximum intensity projections and contrast adjustments, were performed in ImageJ (NIH). Images within each figure were all collected under the same conditions and contrast adjusted identically. For metaphase arrest with Cdc20 depletions, asynchronous cultures in rich media with 1% galactose and 1% raffinose were washed three times and switched to media containing 2% glucose for 2.5 h before imaging. For mitotic chromosome segregation assays, cells were grown overnight in media selective for the lacO cassette and green fluorescent protein-tagged LacI, and then switched to rich media for 5 h. The cells were then fixed with 4% formaldehyde, washed once, stored in storage solution (100 mM potassium phosphate, pH 7.5, 1 M sorbitol) at 4 °C and imaged no more than 2 days later. For meiosis microscopy, saturated cultures in YPD were pelleted, re-suspended in 1% potassium acetate, and incubated with rotation for 24 h at 23 °C. For imaging after microtubule depolymerization, cultures undergoing exponential growth were treated with 10 µg ml^{-1} benomyl and 15 µg ml^{-1} nocodazole for 15 min, washed once with water, put on agar pads with complete synthetic media and 10 µg m l^{-1} benomyl, and immediately imaged.

Tetrad dissection. Diploids were sporulated by transferring patches of yeast from YPD plates grown overnight at 30 °C to sporulation plates (8.2 mg ml^{-1} sodium acetate, 0.35 mg ml^{-1} magnesium sulphate, 1.9 mg ml^{-1} potassium chloride, 1.2 mg ml^{-1} sodium chloride, 16 mg ml^{-1} agar) and incubated at 23 °C for 2–3 days. Spores were then washed with water and digested with 1 mg ml^{-1} zymolyase for 5 min at 30 °C and then dissected onto YPD plates. Genotyping was performed by replica plating colonies onto selective media. For Sli15 constructs integrated at their endogenous loci, a G418 resistance cassette was integrated 1.5 kilobases upstream of either the wild-type or mutant protein for genotyping purposes.

Minichromosome loss assay. Cultures were started overnight in selective (−His) media and then diluted into YPD and grown without selection for 6 h. The cultures were then plated on synthetic media with low (6 µg ml^{-1}) adenine to enhance the colour change[30] and grown for 3 days at 23 °C. The per cent of red or sectored versus completely white colonies was then counted.

29. Longtine, M. S. *et al.* Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* **14,** 953–961 (1998).
30. Hieter, P., Mann, C., Snyder, M. & Davis, R. W. Mitotic stability of yeast chromosomes: a colony color assay that measures nondisjunction and chromosome loss. *Cell* **40,** 381–392 (1985).

doi:10.1038/nature12052

Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX

Myunggon Ko[1]*, Jungeun An[1]*, Hozefa S. Bandukwala[1], Lukas Chavez[1], Tarmo Äijö[1,2], William A. Pastor[1]†, Matthew F. Segal[1], Huiming Li[3]†, Kian Peng Koh[3]†, Harri Lähdesmäki[2], Patrick G. Hogan[1], L. Aravind[4] & Anjana Rao[1,3,5]

TET (ten-eleven-translocation) proteins are Fe(II)- and α-ketoglutarate-dependent dioxygenases[1–3] that modify the methylation status of DNA by successively oxidizing 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxycytosine[1,3–5], potential intermediates in the active erasure of DNA-methylation marks[5,6]. Here we show that IDAX (also known as CXXC4), a reported inhibitor of Wnt signalling[7] that has been implicated in malignant renal cell carcinoma[8] and colonic villous adenoma[9], regulates TET2 protein expression. *IDAX* was originally encoded within an ancestral *TET2* gene that underwent a chromosomal gene inversion during evolution, thus separating the TET2 CXXC domain from the catalytic domain. The IDAX CXXC domain binds DNA sequences containing unmethylated CpG dinucleotides, localizes to promoters and CpG islands in genomic DNA and interacts directly with the catalytic domain of TET2. Unexpectedly, IDAX expression results in caspase activation and TET2 protein downregulation, in a manner that depends on DNA binding through the IDAX CXXC domain, suggesting that IDAX recruits TET2 to DNA before degradation. IDAX depletion prevents TET2 downregulation in differentiating mouse embryonic stem cells, and short hairpin RNA against IDAX increases TET2 protein expression in the human monocytic cell line U937. Notably, we find that the expression and activity of TET3 is also regulated through its CXXC domain. Taken together, these results establish the separate and linked CXXC domains of TET2 and TET3, respectively, as previously unknown regulators of caspase activation and TET enzymatic activity.

TET proteins are restricted to Metazoa and their presence is strictly correlated with the presence of cytosine methylation[2,10]. Most animals have a single TET orthologue, characterized by an amino-terminal CXXC-type zinc finger domain and a carboxy-terminal catalytic Fe(II)- and α-ketoglutarate-dependent dioxygenase domain with an inserted cysteine-rich domain[2,10]. In jawed vertebrates, the *TET* genes underwent triplication, and a subsequent chromosomal inversion split the *TET2* gene into distinct segments encoding the catalytic and CXXC domains[2,10] (Fig. 1a). The ancestral CXXC domain of *TET2* is now encoded by a distinct gene, *IDAX*, which is transcribed in the opposite direction (Fig. 1b and Supplementary Fig. 1a). Given the evolutionary relation between *TET2* and *IDAX*, and the strong sequence conservation of IDAX across species (Supplementary Fig. 1b), we asked whether IDAX could influence the nuclear function of TET2.

We assessed the DNA-binding specificity of the IDAX CXXC domain. A glutathione *S*-transferase-tagged mouse IDAX CXXC-domain fusion protein (GST–IDAX(CXXC)) bound DNA oligonucleotides containing a single unmethylated CpG considerably more efficiently than oligonucleotides containing methylated CpG or no CpG (TpG), as also confirmed by competition with excess unlabelled oligonucleotides; the CFP1 CXXC domain, which preferentially binds DNA sequences containing unmethylated CpG[11], was used as a control (Fig. 1c, left panel and Supplementary Fig. 2). The CXXC domain is found in several proteins involved in DNA methylation and chromatin modification[10–16] (Supplementary Fig. 1c, d). In the structures of the DNMT1–DNA complex and the MLL CXXC domain, short RSKQ and IKKQ loops mediate base-specific contacts with DNA containing unmethylated CpG dinucleotides[12,13,15]. These loops align with the RMKQ motif of KDM2A (also known as CXXC8) required for DNA binding[16], the IRQ sequence of the CFP1 CXXC domain that is critical for recognition of unmethylated CpG[14], and the TGHQ sequence in the IDAX CXXC domain (Supplementary Fig. 1c). We generated a homology model of the IDAX CXXC domain and showed that a TGHQ>AGAA substitution abrogated DNA binding (Fig. 1c, right panel and Supplementary Fig. 3). DNA tethering was required to maintain IDAX in the nucleus, as Myc epitope-tagged wild-type IDAX and TET2 were exclusively nuclear in HEK293T cells, whereas the IDAX DNA-binding mutant (IDAX^{DBM}) was partially present in the cytoplasm (Supplementary Fig. 4).

To determine the genomic distribution of IDAX, we used HEK293T cell lines stably expressing full-length Myc–IDAX or Myc–IDAX^{DBM} (Supplementary Fig. 5a). Chromatin immunoprecipitation followed by next-generation sequencing identified 17,089 peaks for IDAX but only 38 peaks for IDAX^{DBM} (Supplementary Tables 1, 2). Most of the IDAX peaks were located at promoters/transcription start sites (TSSs) and CpG islands (CGIs) of high CpG content (Fig. 1d–g, Supplementary Tables 1, 2 and Supplementary Fig. 5b, c). IDAX-bound regions were enriched for sequences containing CG dinucleotides, tandem cytosines (as suggested previously for the TET3 CXXC domain[17]) or both (Fig. 1h and Supplementary Table 3), and were strongly associated with genes involved in messenger RNA splicing and transcriptional elongation (Supplementary Fig. 5d, e). Thus IDAX preferentially associates with CpG-rich regions containing unmodified cytosines.

To find out whether IDAX and TET2 physically interact, we incubated GST–IDAX(CXXC) (Supplementary Fig. 2a) with lysates from HEK293T cells expressing the TET2 catalytic domain (TET2^{CD}) or the N-terminal region lacking the catalytic domain (TET2$^{\Delta CD}$) (Supplementary Fig. 6a, b). The TET2 catalytic domain bound strongly to both the wild-type IDAX and IDAX^{DBM} CXXC domain (Fig. 2a), whereas TET2$^{\Delta CD}$ bound more weakly (Supplementary Fig. 6c). Co-immunoprecipitation assays confirmed a strong interaction between Myc–IDAX^{DBM} (which does not downregulate TET2 protein; see below) and full-length TET2 or TET2^{CD} in HEK293T cells (Fig. 2b). The interaction was direct, as Flag-tagged TET2^{CD} expressed in insect cells bound GST–IDAX(CXXC), but not GST–CFP1(CXXC), in pull-down assays (Supplementary Fig. 6d, e).

[1]Division of Signaling and Gene Expression, La Jolla Institute for Allergy & Immunology, La Jolla, California 92037, USA. [2]Department of Information and Computer Science, Aalto University School of Science, FI-00076 Aalto, Finland. [3]Harvard Medical School and Program in Cellular and Molecular Medicine, Children's Hospital, Boston, Massachusetts 02115, USA. [4]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. [5]Sanford Consortium for Regenerative Medicine, La Jolla, California 92037, USA. †Present addresses: Department of Molecular, Cell and Developmental Biology, University of California at Los Angeles, Terasaki Life Sciences Building, 610 Charles Young Drive East, Los Angeles, California 90095-723905, USA (W.A.P.); Bristol-Myers Squibb, 700 Bay Road, Redwood City, California 94063, USA (H.Li); KU Leuven Department of Development and Regeneration & Stem Cell Institute Leuven, Herestraat 49, 3000 Leuven, Belgium (K.P.K.).
*These authors contributed equally to this work.

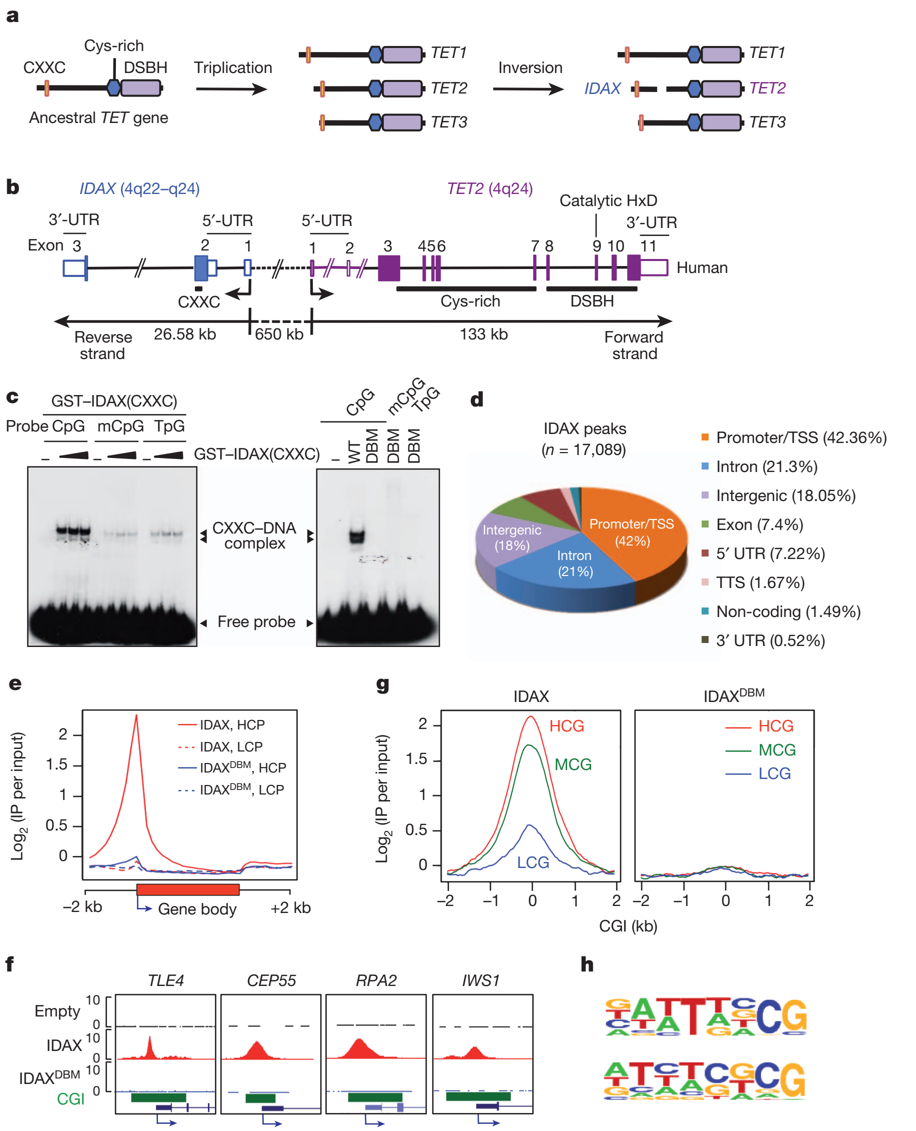


Figure 1 | IDAX preferentially binds CpG-rich DNA. **a**, The evolutionary origin of *IDAX*. Cys-rich, cysteine-rich; DSBH, double-stranded β-helix. **b**, The human *IDAX* and *TET2* genes are transcribed in opposite directions. UTR, untranslated region. **c**, Left, GST–IDAX(CXXC) preferentially binds unmethylated CpGs in DNA. Right, the TGHQ > AGAA substitution abrogates the DNA-binding activity of the IDAX CXXC domain. mCpG, methylated CpG; WT, wild type. **d**, Genome-wide distribution of IDAX binding sites in HEK293T cells. TTS, transcription termination site. **e**, IDAX binds preferentially to TSSs at high-CpG-density (HCP) but not low-CpG-density (LCP) promoters. IP, immunoprecipitate. **f**, Examples of IDAX peaks at CGI promoters. **g**, Distribution of IDAX across high (HCG)-, medium (MCG)- and low (LCG)-CpG-density CGIs. **h**, DNA motifs conserved in IDAX-bound loci revealed by *de novo* motif discovery analysis.

Unexpectedly, transient co-expression of Myc–IDAX and Flag–haemagglutinin-tagged TET2 in HEK293T cells led to the disappearance of TET2 protein and a decrease in 5-hydroxymethylcytosine (5hmC), with only a minor effect on *TET2* mRNA (Fig. 2c and Supplementary Fig. 7). IDAX DNA-binding activity was required, as co-expressed Myc–IDAX^{DBM} did not decrease TET2 protein or 5hmC (Fig. 2d, e and Supplementary Fig. 8). Myc–IDAX^{DBM} was expressed at considerably higher levels than wild-type Myc–IDAX (Fig. 2d, e, g and Supplementary Fig. 8), suggesting that DNA-bound IDAX recruits a degradation complex that targets both IDAX and TET2 (see below, Supplementary Fig. 16). Treatment of cells co-expressing Myc–IDAX and Flag–HA–TET2 with proteasome inhibitors variably rescued the loss of TET2 protein, whereas treatment with lysosomal inhibitors had no effect (Supplementary Fig. 9a, b). However, IDAX was unable to decrease Myc–TET2 protein levels in cells treated with the pan-caspase inhibitor Z-VAD-FMK (Fig. 2f); moreover, wild-type IDAX induced nuclear cleavage of poly-(ADP-ribose) polymerase (PARP), a marker for caspase activation, whereas IDAX^{DBM} did not (Fig. 2g and Supplementary Fig. 9c). TET2 was a direct target for caspase cleavage, as shown by treatment of HEK293T cell lysates containing Myc–TET2 with recombinant active human caspase 3 and caspase 8 (Fig. 2h and Supplementary Fig. 9d, e). Neither wild-type IDAX nor IDAX^{DBM} influenced the enzymatic activity of TET2 *in vitro* (Supplementary Fig. 10), indicating that the loss of genomic 5hmC in cells co-expressing TET2 and IDAX reflects the loss of TET2 protein rather than any direct interference with TET2 enzymatic activity.

Regulation of TET2 by IDAX was observed in three independent systems. *Idax* mRNA levels were low in murine v6.5 embryonic stem cells (ESCs), but increased progressively upon leukaemia inhibitory factor (LIF) withdrawal and supplementation of the culture medium with retinoic acid (Fig. 3a, left panel), concomitantly with decreased and increased expression of *Tet1* and *Tet3*, respectively[18] (Supplementary Fig. 11a). Under these conditions, *Tet2* mRNA levels were only slightly altered (Fig. 3a, right panel), but TET2 protein levels decreased precipitously over a 3-day period (Fig. 3b; for TET2 antibody specificity see Supplementary Fig. 11b–d). Lentiviral transduction of v6.5 ESCs with three effective short hairpin RNAs (shRNAs) against *Idax* (*shIdax* no. 1, 3 and 4; Fig. 3c) substantially protected against the differentiation-induced downregulation of TET2 protein, whereas transduction with an ineffective shRNA, *shIdax* no. 2, did not (Fig. 3d). Thus in differentiating murine ESCs, TET2 protein downregulation can be directly attributed to IDAX. In addition, transduction of the human U937 monocytic cell line, which barely expresses TET2, with four separate lentiviral shRNAs against *IDAX* resulted in strong TET2 protein expression (Fig. 3e), suggesting that endogenous IDAX maintains low endogenous TET2 levels in U937 cells. Finally, IDAX expression, like TET2 deficiency[3,19], skewed the differentiation of murine bone marrow haematopoietic

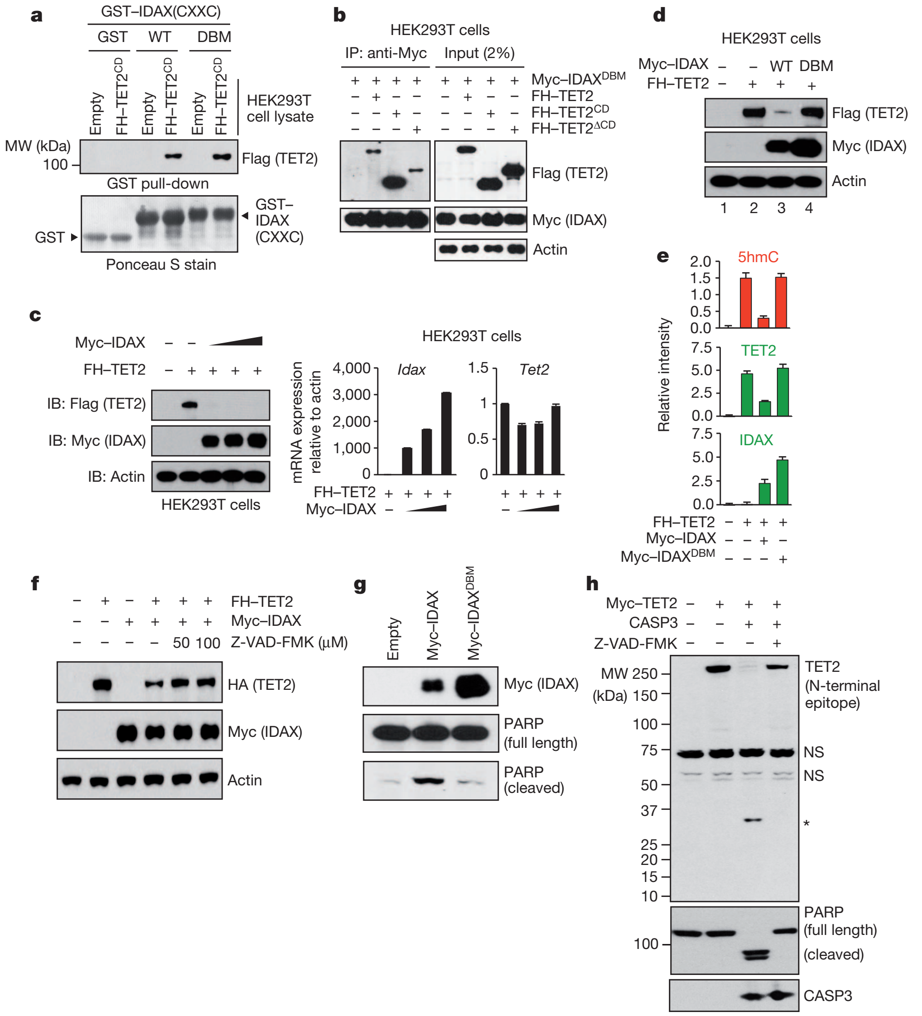


Figure 2 | IDAX downregulates TET2 protein through caspase activation. **a**, Both wild-type IDAX and IDAX^{DBM} CXXC domains physically associate with TET2. FH, Flag–HA tag. **b**, Co-immunoprecipitation of TET2 with IDAX^{DBM} in HEK293T cell lysates. **c**, IDAX expression decreases the expression of TET2 protein (left) but not *TET2* mRNA (right). IB, immunoblot. **d**, The DNA-binding activity of IDAX is required to cause loss of TET2 protein. **e**, IDAX^{DBM} does not affect TET2 protein or 5hmC levels in HEK293T cells. Error bars indicate s.d. ($n = 24$ wells). **f**, IDAX does not decrease TET2 protein levels in the presence of pan-caspase inhibitor Z-VAD-FMK. **g**, PARP cleavage induced by IDAX but not IDAX^{DBM} in transfected HEK293T cells. **h**, TET2 is a substrate for recombinant active caspase 3 *in vitro*. MW, molecular weight. NS, nonspecific.

stem/progenitor cells towards the monocyte/macrophage lineage (Supplementary Fig. 12).

The direct binding of the CXXC domain of IDAX to the catalytic domain of TET2 suggested the possibility of corresponding interactions between the linked CXXC and catalytic domains of TET1 and TET3. Indeed, a CXXC domain mutant of TET3 (80-THQ-82 > 80-AAA-82, TET3 CXXC^{mut}, Fig. 4a) resembled IDAX^{DBM} in several respects: it was expressed at higher levels in HEK293T cells compared with wild-type or catalytically inactive TET3 (Fig. 4b, c), a fraction of TET3 CXXC^{mut} was aberrantly present in the cytoplasm (Fig. 4d, top panel) and TET3 CXXC^{mut} was weaker than wild-type TET3 at inducing PARP cleavage in the nucleus (Fig. 4d, middle panel). Cells expressing TET3 CXXC^{mut} displayed higher levels of genomic 5hmC compared with cells expressing wild-type TET3 (Fig. 4e). This could not be attributed simply to increased protein levels of TET3 CXXC^{mut} versus wild-type TET3, because the nuclear levels of these two proteins were comparable (Fig. 4d, top panel). Together, these results indicate that TET3 autoregulates its protein levels via caspase activation, through a mechanism that requires DNA binding by the CXXC domain. The catalytic activity of TET3 may be controlled by an intramolecular interaction involving the CXXC domain (Fig. 4g); alternatively, the increased 5hmC in cells expressing TET3 CXXC^{mut} could be due to loss of DNA tethering such that the mutant protein is more freely distributed than wild-type TET3 in the nucleus. By contrast, the expression and activity of TET1 CXXC^{mut} were similar to those of wild-type TET1 (Supplementary Fig. 13), suggesting that TET1 and TET3 are regulated by different mechanisms.

The IDAX-related protein CXXC5 (also known as RINF) is over-expressed in solid cancers and its high expression correlates with poor prognosis in breast cancer[20]. Notably, CXXC5 resembles IDAX in being a negative regulator of Wnt signalling[21]; moreover, like TET2, CXXC5 has a role in normal myelopoiesis as well as in myeloid cancers and the *CXXC5* gene is often deleted in myeloid leukaemias[22]. Like IDAX, mouse CXXC5 downregulated TET2 protein expression in a dose-dependent manner (Fig. 4f and Supplementary Fig. 14), suggesting that CXXC5 might also regulate the genomic localization, expression and enzymatic function of TET2.

We present a striking example of how a gene fission event, followed by a chromosomal inversion, facilitated the emergence of a previously unknown regulatory mechanism for DNA cytosine modification in vertebrates. The IDAX CXXC domain, which separated from the TET2 catalytic domain during vertebrate evolution[2,10], binds TET2 and potentially recruits it to promoters and CpG islands in genomic DNA. We propose that DNA-bound IDAX activates one or more caspases, perhaps at a transcriptional level through promoter/TSS binding (Supplementary Fig. 15). Indeed, caspase 3 activation and IDAX upregulation occur with similar kinetics during ESC differentiation[23]. As proposed for ubiquitylation[24], caspase-mediated cleavage could be involved in the initial activation as well as the subsequent degradation of TET2 and IDAX (Supplementary Fig. 16).

The cellular and subcellular distribution of IDAX, and its sensitivity to local environmental signals, may contribute to differences in 5hmC and TET2 levels in different cell types and organs[3,25–27]. IDAX^{DBM} is partly in the cytoplasm, and IDAX is reported to bind dishevelled in the cytoplasm in the absence of Wnt signalling[7]. Because IDAX is expressed at very low levels in ESCs but becomes upregulated upon differentiation, the genomic distribution of TET2 and oxidized methylcytosines in ESCs would be predicted to differ from those in

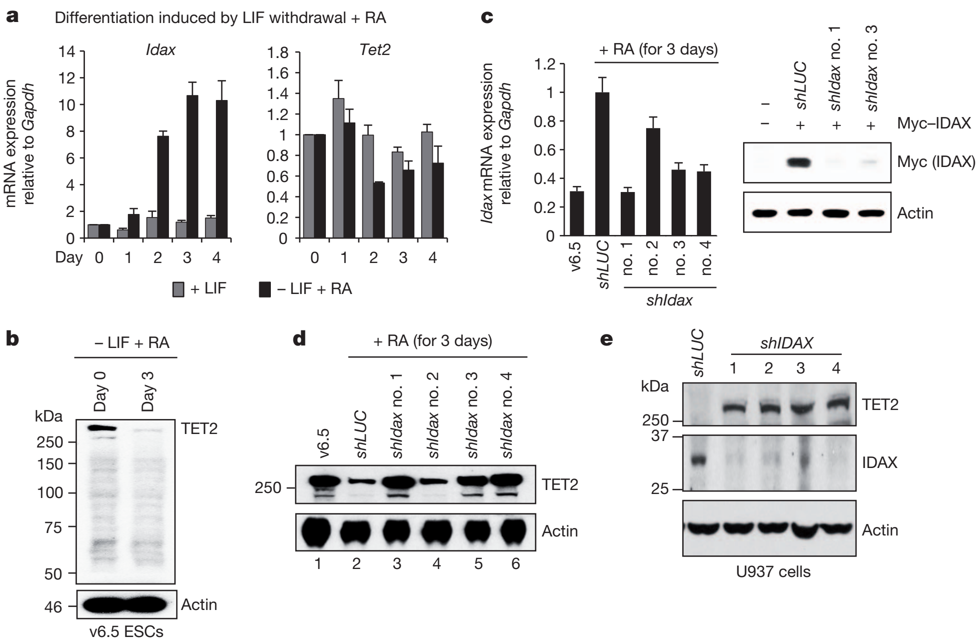


Figure 3 | Reciprocal relationship between TET2 and IDAX in mouse ESCs and human U937 cells. a, *Tet2* and *Idax* mRNA levels in v6.5 ESCs differentiated by LIF withdrawal plus 1 μM retinoic acid (− LIF + RA). Mean ± s.e.m. of 4–6 independent experiments is shown. **b**, TET2 protein expression decreases within 3 days of ESC differentiation. **c**, *shIdax* no. 1, 3 and 4 effectively deplete *Idax* mRNA in differentiating mouse ESCs (left), and shRNAs no. 1 and 3 decrease Myc–IDAX protein levels in human HEK293T cells transfected with complementary DNA lacking the 3′ UTR (right). *shIdax* no. 4 was not tested in HEK293T cells as it is directed against the *Idax* 3′ UTR. Error bars show the range of duplicates. LUC, luciferase. **d**, IDAX depletion prevents the decrease in TET2 protein expression in differentiating ESCs. Undifferentiated ESCs (lane 1) served as a control. **e**, Depletion of endogenous IDAX in the human monocytic cell line U937 augments expression of endogenous TET2 protein.

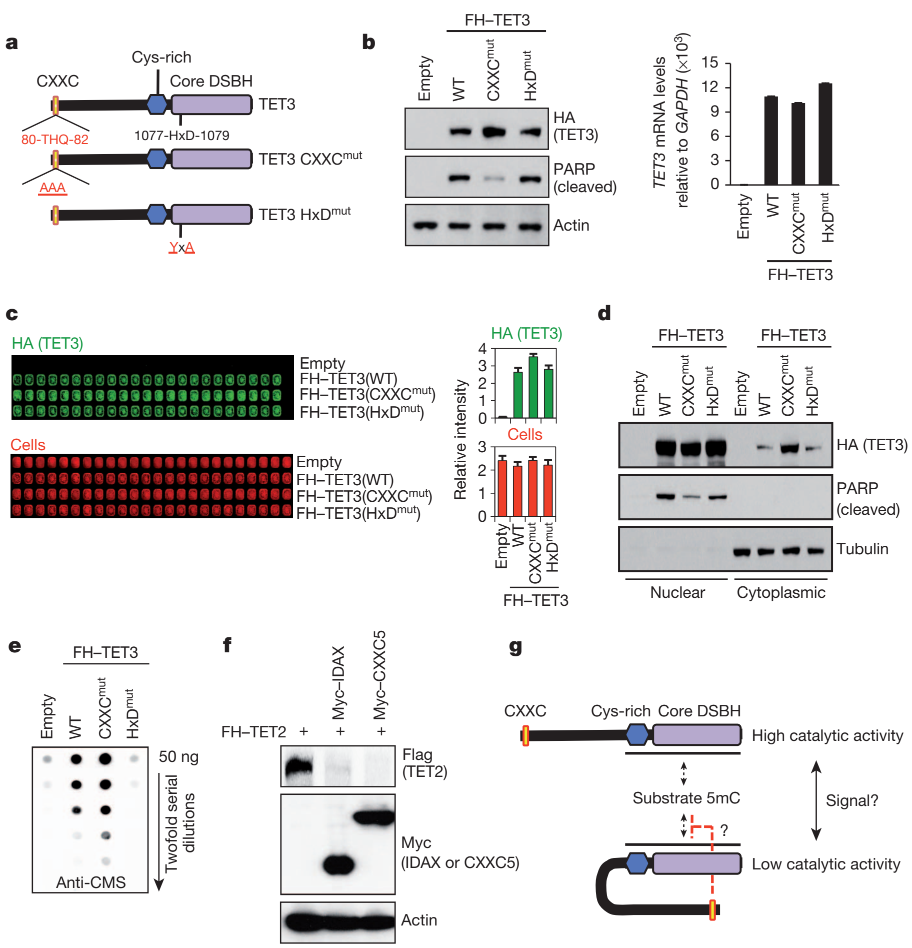


Figure 4 | Negative regulation of TET3 by its CXXC domain. a, Schematic representation of TET3, TET3 CXXC^{mut} and catalytically inactive TET3 HxD^{mut}. **b**, Increased protein expression of TET3 CXXC^{mut} relative to wild-type TET3 or TET3 HxD^{mut} (left), without a change in *TET3* mRNA levels (right). **c**, In-cell western assays confirm that TET3 CXXC^{mut} is expressed at higher levels than wild-type TET3 or TET3 HxD^{mut}. **d**, Wild-type TET3 and TET3 HxD^{mut} are mainly nuclear, whereas TET3 CXXC^{mut} is found in both cytoplasmic and nuclear fractions and is less effective at inducing PARP cleavage than wild-type TET3. **e**, Cells expressing TET3 CXXC^{mut} show higher genomic 5hmC than cells expressing wild-type TET3 or TET3 HxD^{mut} (anti-cytosine 5-methylenesulphonate (CMS) dot blot[3]). **f**, CXXC5 expression results in a decrease in protein levels of co-expressed TET2 in HEK293T cells. **g**, Potential intramolecular (autoinhibitory) interaction between the linked CXXC and catalytic domains of TET3. 5mC, 5-methylcytosine.

more differentiated tissues, such as the primitive streak of the mouse embryo, in which Wnt signalling is prominent[28]. Further analyses will resolve these questions.

Many malignant tissues have been reported to contain low levels of 5hmC compared with normal tissues[29,30], and *TET2* loss of function correlates with decreased 5hmC levels in patients with myeloid malignancies[3]. However, an appreciable proportion of patients with wild-type *TET2* show low 5hmC[3], suggesting the existence of additional factors that impair TET2 expression at the protein level and/or TET2 enzymatic activity. For instance, *IDAX* overexpression, reported in villous adenomas of the colon[9], might promote TET2 degradation, leading to depletion of 5hmC in TET2 target genes. Conversely, the homozygous deletion of *IDAX* reported in an aggressive renal cell carcinoma[8] could result in TET2 overexpression and its aberrant recruitment to genomic regions where it is not normally found. It will be important in future studies to explore the genomic targets of TET2, IDAX and CXXC5 in normal development and cancer, and to define the effects of cancer-associated mutations in these proteins on patterns of cytosine modification in DNA.

METHODS SUMMARY

ChIP-seq and data analysis. Chromatin was prepared from two biological replicates of HEK293T cells expressing empty vector, Myc–IDAX or Myc–IDAX^{DBM} using truChIP High Cell Chromatin Shearing Kit with Non-ionic Shearing Buffer (Covaris), sheared into 200–300 base-pair fragments using a Covaris S2 instrument as per manufacturer's instructions, and immunoprecipitated using anti-Myc conjugated to agarose (Sigma). Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) libraries were constructed using TruSeq Sample Prep kit (Illumina) and single-end sequencing was performed using Illumina HiSeq 2000. Sequencing reads were mapped against hg19 using Bowtie. Mapping results from both ChIP-seq experiments were combined and IDAX-enriched regions were identified by MACS peak-calling software.

Cell cultures and lentiviral knockdown. For *Idax* and *IDAX* knockdown, the human monocytic U937 cell line or mouse v6.5 ESC cell line (cultured as described[18]) were transduced with pLKO.1_Puro lentivirus expressing shRNA against *Idax* and *IDAX*, respectively (see Supplementary Table 4 for target sequences). Transduced cells were selected with puromycin. Lentivirus expressing shRNA against luciferase was used as a control.

Full Methods and any associated references are available in the online version of the paper.

Received 15 February 2012; accepted 28 February 2013.
Published online 7 April 2013.

1. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324,** 930–935 (2009).
2. Iyer, L. M., Tahiliani, M., Rao, A. & Aravind, L. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle* **8,** 1698–1710 (2009).
3. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant *TET2*. *Nature* **468,** 839–843 (2010).
4. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333,** 1300–1303 (2011).
5. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–1307 (2011).
6. Maiti, A. & Drohat, A. C. Dependence of substrate binding and catalysis on pH, ionic strength, and temperature for thymine DNA glycosylase: Insights into recognition and processing of G.T mispairs. *DNA Repair* **10,** 545–553 (2011).
7. Hino, S. *et al.* Inhibition of the Wnt signaling pathway by Idax, a novel Dvl-binding protein. *Mol. Cell. Biol.* **21,** 330–342 (2001).
8. Kojima, T. *et al.* Decreased expression of CXXC4 promotes a malignant phenotype in renal cell carcinoma by activating Wnt signaling. *Oncogene* **28,** 297–305 (2009).
9. Nguyen, A. V., Albers, C. G. & Holcombe, R. F. Differentiation of tubular and villous adenomas based on Wnt pathway-related gene expression profiles. *Int. J. Mol. Med.* **26,** 121–125 (2010).
10. Iyer, L. M., Abhiman, S. & Aravind, L. Natural history of eukaryotic DNA methylation systems. *Prog. Mol. Biol. Transl. Sci.* **101,** 25–104 (2011).
11. Lee, J. H., Voo, K. S. & Skalnik, D. G. Identification and characterization of the DNA binding domain of CpG-binding protein. *J. Biol. Chem.* **276,** 44669–44676 (2001).
12. Cierpicki, T. *et al.* Structure of the MLL CXXC domain–DNA complex and its functional role in MLL-AF9 leukemia. *Nature Struct. Mol. Biol.* **17,** 62–68 (2010).
13. Allen, M. D. *et al.* Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. *EMBO J.* **25,** 4503–4512 (2006).
14. Xu, C., Bian, C., Lam, R., Dong, A. & Min, J. The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nature Commun.* **2,** 227 (2011).
15. Song, J., Rechkoblit, O., Bestor, T. H. & Patel, D. J. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science* **331,** 1036–1040 (2011).
16. Blackledge, N. P. *et al.* CpG islands recruit a histone H3 lysine 36 demethylase. *Mol. Cell* **38,** 179–190 (2010).
17. Xu, Y. *et al.* Tet3 CXXC domain and dioxygenase activity cooperatively regulate key genes for *Xenopus* eye and neural development. *Cell* **151,** 1200–1213 (2012).
18. Koh, K. P. *et al.* Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell* **8,** 200–213 (2011).
19. Ko, M. *et al.* Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proc. Natl Acad. Sci. USA* **108,** 14566–14571 (2011).
20. Knappskog, S. *et al.* RINF (CXXC5) is overexpressed in solid tumors and is an unfavorable prognostic factor in breast cancer. *Ann. Oncol.* **22,** 2208–2215 (2011).
21. Kim, M. S. *et al.* A novel Wilms tumor 1 (WT1) target gene negatively regulates the WNT signaling pathway. *J. Biol. Chem.* **285,** 14585–14593 (2010).
22. Pendino, F. *et al.* Functional involvement of RINF, retinoid-inducible nuclear factor (CXXC5), in normal and tumoral human myelopoiesis. *Blood* **113,** 3172–3181 (2009).
23. Fujita, J. *et al.* Caspase activity mediates the differentiation of embryonic stem cells. *Cell Stem Cell* **2,** 595–601 (2008).
24. Geng, F., Wenzel, S. & Tansey, W. P. Ubiquitin and proteasomes in transcription. *Annu. Rev. Biochem.* **81,** 177–201 (2012).
25. Nestor, C. E. *et al.* Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Res.* **22,** 467–477 (2012).
26. Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS ONE* **5,** e15367 (2010).
27. Haffner, M. C. *et al.* Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget* **2,** 627–637 (2011).
28. Sokol, S. Y. Maintaining embryonic stem cell pluripotency with Wnt signaling. *Development* **138,** 4341–4350 (2011).
29. Yang, H. *et al.* Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene* **32,** 663–669 (2013).
30. Lian, C. G. *et al.* Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell* **150,** 1135–1146 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. Seumois, M. Ku and J. Day for help with library preparation, B. Ren for use of his Illumina Hi-Seq 2000, J. A. Zepeda-Martínez for the recombinant Flag–TET2^{CD}, and members of the Rao laboratory for discussions. This work was supported by National Institutes of Health (NIH) R01 grants HD065812 and CA151535, grant RM-01729 from the California Institute of Regenerative Medicine and Translational Research, grant TRP 6187-12 from the Leukemia and Lymphoma Society (to A.R.) and NIH R01 grant AI40127 (to P.G.H. and A.R). We also gratefully acknowledge a Special Fellow Award from the Leukemia and Lymphoma Society (to M.K.), postdoctoral fellowships from the Lady Tata Memorial Trust and from the GlaxoSmithKline-Immune Disease Institute Alliance (to H.S.B.) and a predoctoral graduate research fellowship from the National Science Foundation (to W.A.P.).

Author Contributions L.A., P.G.H. and A.R. conceived the project and supervised project planning and execution. M.K. and J.A. performed cellular and molecular experiments including ChIP-seq, gene knockdown, establishment of stable cell lines, site-directed mutagenesis, dot blot, immunocytochemistry, *in vitro* caspase and TET assays, and *in vitro* differentiation studies. J.A. performed the in-cell western blots. H.S.B. obtained the initial data showing downregulation of TET2 protein by IDAX. M.K. conducted the electrophoretic mobility shift assays with help from W.A.P. and M.F.S. H.Li and P.G.H. generated the homology model of the IDAX CXXC domain. K.P.K. provided mRNAs from ESC samples. L.C., T.A. and H.Lähdesmäki performed the bioinformatic analyses of ChIP-seq data. M.K. and A.R. wrote the manuscript with input from other authors.

Author Information The ChIP-seq data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE42958. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.R. (arao@liai.org).

METHODS

Recombinant protein expression and purification. Competent BL21 cells were transformed with appropriate pGEX constructs, then induced to produce proteins using 100 ml Overnight Express Autoinduction System 2 (EMD Biosciences). Induced bacteria were collected and washed with 1× PBS. Next, cells were re-suspended in 15 ml PBS containing protease inhibitors (Roche) and lysed by sonication. The lysate was incubated in the presence of lysozyme (1 mg ml^{-1}) and Triton X-100 (1%) with gentle rotation for 30 min at 4 °C and hard spun at 20,000*g* for 30 min at 4 °C. The soluble fraction was incubated with 2 ml Glutathione Sepharose 4B slurry (HE Healthcare) with gentle rotation for 30 min at 4 °C and washed with cold 1× PBS three to four times. The recombinant GST fusion proteins were incubated in elution buffer (50 mM Tris-HCl, pH 8.0, 10 mM reduced glutathione) for 10 min at 4 °C and the eluted proteins were collected by centrifugation. The elution and incubation steps were repeated three times to collect all proteins. Subsequently, the eluted protein was dialysed overnight in cold 1× PBS, 10% glycerol, 10 μM $ZnCl_2$ and 10 mM β-mercaptoethanol and flash frozen. Protein was concentrated by Amicon Ultra Spin columns 10,000 molecular weight cutoff. Protein expression and purification in Sf9 cells were described previously[1]. Integrity and purity of proteins were verified by SDS–PAGE and Coomassie blue staining, and purified protein was quantified using a Bradford assay.

Electrophoretic mobility shift assay (EMSA). Single-stranded oligonucleotides used for substrate preparation were synthesized at IDT (Sequences in Supplementary Fig. 2b). One nanomole of complementary single-stranded oligonucleotides was mixed in annealing buffer (10 mM Tris, pH 8.0, 50 mM NaCl, 1 mM EDTA), boiled for 4 min and gradually cooled overnight by transferring to a water bath preheated to 90 °C and turned off. The annealed oligonucleotides and single-stranded oligonucleotides as controls were separated by running on a 15% acrylamide gel in 0.5× Tris-borate-EDTA (TBE) gel. The gel was placed on an X-ray intensifying screen and irradiated with short-wave ultraviolet light to visualize DNA and excise the region of the gel containing the correct sized double-stranded oligonucleotides. The excised gel slice was crushed and incubated with 10 mM Tris, pH 8.0, 0.1 mM EDTA and 300 mM sodium acetate, pH 7.4, overnight at 37 °C with agitation. Supernatant was extracted, spun for 5 min at 14,000 r.p.m. to remove undissolved acrylamide, and the eluted DNA was purified by ethanol precipitation. A total of 100 ng double-stranded DNA was radiolabelled using T4 PNK enzyme (NEB) according to the manufacturer's protocols. Unincorporated ATP was removed using an illustra G-25 Micro Column (GE Healthcare). For the experiment shown in the left panel of Fig. 1c, 0, 100, 200 and 400 ng of GST–IDAX CXXC domain was used. Except where otherwise noted, 500 ng of purified GST-tagged wild-type IDAX or IDAX^{DBM} CXXC domain was incubated in a total of 20 μl reaction mixture containing 1 ng of radiolabelled oligonucleotide and 2 μg of poly(dA:dT), 20 mM HEPES, pH 7.9, 40 mM KCl, 2.5 mM $MgCl_2$, 1 mM dithiothreitol (DTT) and 5% glycerol for 10 min on ice, then an additional 30 min at room temperature (~25 °C). For competition assays, the extracts were pre-incubated with unlabelled DNA competitors before addition of the reaction mixture. The resulting protein–DNA complexes were resolved in a non-denaturing 5% acrylamide gel in 0.5× TBE buffer. Gels were dried and visualized by autoradiography.

Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq). Chromatin was sheared using truChIP High Cell Chromatin Shearing Kit with Non-ionic Shearing Buffer (Covaris). In brief, cells were cross-linked with 1× Covaris fixing buffer at room temperature for 5 min with rotation, followed by addition of 1× Covaris quenching buffer. After further incubation at room temperature for 5 min with rotation, cells were washed with PBS twice at 4 °C and nuclei were prepared as per manufacturer's instructions. The nuclei pellet was re-suspended in 1× Covaris shearing buffer and chromatin was sheared into 200–300 base-pair (bp) fragments using a Covaris S2 instrument (duty cycle, 5%; intensity, 4; 200 cycles per burst; 20 min). Chromatin fragments from two biologically independent cells expressing empty vector, Myc–IDAX or Myc–IDAX^{DBM} were immuneprecipitated using anti-Myc antibody conjugated to agarose (Sigma). The ChIP-seq library was constructed using TruSeq Sample Prep kit (Illumina) and Illumina HiSeq sequencing was performed as per manufacturer's instructions.

Data analyses. Sequencing reads were mapped against hg19 using Bowtie[31] (bowtie-0.12.7 -n 2 -l 28 -m 3 -k 1–best–chunkmbs 512 -p 8 -S -q) by suppressing all alignments for a particular read, if more than three reportable alignments exist, and by reporting only the best valid alignment per read. Mapping results from both ChIP-seq experiments per condition were combined and clonal reads (that is, reads mapped to the same genomic position) were replaced by one representative, resulting in 52.4 million, 43 million and 29.9 million reported reads for Myc–IDAX cells, Myc–IDAX^{DBM} cells and cells expressing empty vector, respectively. Genomic regions enriched for short reads obtained from Myc–IDAX or Myc–IDAX^{DBM}, respectively, were identified by using a local background estimated from short reads obtained from cells expressing empty vector using the statistical software MACS[32]. Annotation (hg19) of the obtained peaks was performed by using the annotatePeaks.pl function of the Homer software package[33]. The promoter region was defined as −1 kilobase (kb) to +0.1 kb relative to TSSs. In addition, we created Myc–IDAX- and Myc–IDAX^{DBM}-specific peak sets by excluding all overlapping peaks. For the Myc–IDAX-specific peaks, we performed *de novo* motif discovery by using the findMotifsGenome.pl function of the Homer software package[33] using default parameter settings. Moreover, we calculated the CpG densities of the DNA sequences underlying the identified peaks, by the CpG observed/expected ratio ((number of CpG/(number of × number of G)) × total number of nucleotides in the sequence)[34].

In order to calculate immunoprecipitate enrichment at gene regions and at CpG islands, we accessed hg19 RefSeq gene annotations and CpG island annotations provided by the University of California Santa Cruz (UCSC) genome browser database (11/19/2012)[35]. According to their promoter CpG density, we divided all genes into low-CpG-density promoter (LCP) and high-CpG-density promoter (HCP) genes. For this, we calculated the CpG observed/expected ratio[34] of the sequences underlying the −1 kb to +0.5 kb region around their TSSs. By visual inspection of the CpG density distribution, we decided on a threshold of 0.44 for dividing the genes into LCP and HCP. In order to divide the CpG islands into high CpG (HCG)-, intermediate CpG (ICG)- and low CpG (LCG)-density islands, we calculated the CpG observed/expected ratio[34] of the underlying sequences. Subsequently, we sorted all CpG islands into five equally sized groups according to their descending CpG observed/expected ratio. We used the first, third and fifth group for defining HCG, ICG and LCG islands.

For the gene-centric enrichment analysis, each gene was divided into 20 bins. In addition, the 2-kb region upstream of the TSS, as well as the 2-kb region downstream of the transcription end site, was divided into 10 bins each. For the CpG island and TSS-centric analyses, we divided the −2 kb to +2 kb range around the midpoint of each island (or the TSS, respectively) into 80 bins. Short reads were extended to a length of 200 bp along their sequencing direction and bin-wise coverage was calculated using the Bioconductor[36] environment. In order to avoid division by zero, we added one to each bin of the immunoprecipitate and of the control sample. For each immunoprecipitate sample, a correction value for the library size was estimated by calculating the mean over the ratios of immunoprecipitate over control short-read coverage at genome-wide 100-bp windows. Subsequently, immunoprecipitate signals at the tested bins were divided by the correction value. Enrichment is defined by the $\log_2$ of corrected immunoprecipitate signals divided by the control signals, where Myc–IDAX or Myc–IDAX^{DBM}, respectively, is considered as immunoprecipitate and cells expressing empty vector is considered as control. For the average enrichment profiles, we calculate the $\log_2$ of the mean over all corrected immunoprecipitate over control ratios at every tested position of stacked annotations.

In-cell western blot. In-cell western was performed using standard immunocytochemistry procedures. Diluted DNA was mixed with 0.25 μl lipofectamine 2000 (Invitrogen) diluted in Opti-MEM reduced serum medium (Invitrogen) and incubated at room temperature for 20 min. HEK239T cells were trypsinized, washed and counted, and 2×10^4 cells were plated in each well of amine-coated BD PureCoat 384-well plate (BD Bioscience), followed by addition of the DNA–lipofectamine mixture. Three days later, cells were fixed with 4% paraformaldehyde in PBS for 15 min and permeabilized with 0.2% Triton X-100 in PBS for 15 min at room temperature. Subsequently, DNAs were denatured with 2N HCl at room temperature for 30 min and neutralized with 100 mM Tris-HCl buffer, pH 8.5, for 10 min. Cells were incubated in Odyssey Blocking Buffer (LI-COR Biosciences) diluted 1:1 with PBS at room temperature for 1 h. Next, rabbit anti-5hmC polyclonal antibody (produced in-house; diluted to 1:2,500) for 5hmC staining or mouse anti-HA antibody (Covance, HA.11 clone 16B12, diluted to 1:1,000) and mouse anti-Myc antibody (Sigma, diluted to 1:2,000) for HA–TET2 and Myc–IDAX staining, respectively, were added into the blocking buffer for 3 h at room temperature. The cells were rinsed three times with 0.2% Triton X-100 in PBS and incubated at room temperature for 1 h with an IRDye 800CW-conjugated donkey anti-mouse IgG (1:4,000) or IRDye 680LT-conjugated donkey anti-rabbit IgG secondary antibody (1:1,000) (LI-COR Biosciences) in blocking buffer, and again washed four times with wash buffer, followed by addition of 50 μl PBS. Plates were scanned on the Odyssey Sa Infrared Imaging System and staining was quantified using the scanner software. To quantify cell density, cells were stained with IRDye 680LT maleimide (diluted into 10 ng ml^{-1} in PBS, LI-COR Biosciences) at room temperature for 15 min. After extensive washing with PBS three times, the plate was scanned again and staining was quantified using the scanner software.

Western blotting and caspase-mediated TET2 cleavage *in vitro*. Cells were lysed with radioimmunoprecipitation assay (RIPA) buffer (150 mM NaCl, 50 mM Tris-HCl, pH 8.0, 1% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS) supplemented with protease inhibitor cocktail (Roche) or mixture of protease/phosphatase

inhibitors (20 mM β-glycerophosphate, 10 mM sodium pyrophosphate, 1 mM sodium o-vanadate, 10 μM leupeptin, 10 μg ml^{-1} aprotinin, 1 mM freshly prepared phenylmethylsulphonyl fluoride (PMSF)) and incubated on ice for 20 min. Cell debris was removed by centrifuging at 12,000 r.p.m. for 15 min at 4 °C. The protein concentration was measured by Bradford protein assay. Samples were mixed with SDS sample buffer and boiled for 4 min. Whole-cell lysates were separated by 7.5% or 10% SDS–PAGE and transferred onto nitrocellulose membranes. Proteins were detected by immunoblotting in TBST (150 mM NaCl, 10 mM Tris-Cl, pH 8.0, 0.5% Tween-20) containing 5% low-fat milk and antibodies against TET2 (Abcam, ab94580 or Abiocode, R1086-vp), DDK (Flag) epitope (Origene, TA50011-100), Myc epitope (clone 9E10 or hybridoma supernatant prepared in-house), HA epitope (Covance, HA.11 clone 16B12), actin (Sigma A5441), CASP3, CASP8, PARP and cleaved PARP (Cell Signaling) proteins, followed by incubation with horseradish peroxidase-conjugated secondary antibodies and enhanced chemiluminescence. Except where otherwise noted, immunoblot analyses were performed 48 h after transfection. For *in vitro* TET2 cleavage assay using recombinant caspases, lysates (~20 μg) from HEK293T cells expressing Myc–TET2 proteins were incubated with ~ 0.4 μg of recombinant active human caspases 3 or 8 (BD Biosciences) in 1× caspase 3 (20 mM HEPES, pH 7.4, 0.1% CHAPS, 5 mM DTT, 2 mM EDTA) or caspase 8 (20 mM HEPES, pH 7.4, 0.1% CHAPS, 5 mM DTT and 2 mM EDTA, 5% sucrose) assay buffer supplemented with protease/phosphatase inhibitors, respectively at 37 °C for 3 h with gentle shaking. As a control, Z-VAD-FMK (100 μM) was added to the reaction mixture before addition of caspases. Lysates were then separated by SDS–PAGE.

ESC culture and transduction. v6.5 mouse ESCs were maintained in culture as previously described[18]. In brief, v6.5 mouse ESCs were cultured on primary mouse embryonic fibroblasts that were pre-treated with mitomycin C (Sigma). The culture medium contains DMEM knockout (Invitrogen), 15% Stasis Stem Cell FBS (Gemini Bio-Products), 0.1 mM non-essential amino acids (Invitrogen), 2 mM L-glutamine (Invitrogen), 0.1 mM β-mercaptoethanol (Invitrogen), 50 U ml^{-1} penicillin/streptomycin (Invitrogen) and 0.2% LIF-conditioned medium (produced in-house). Culture media were changed every day with fresh complete ESC media. In experiments to induce differentiation, cells were first trypsinized (or split using TrypLE Express) and plated onto gelatin-coated plates for 45 min two times to remove feeder cells. The floating feeder-depleted v6.5 cells were collected and washed twice with 1× PBS. Next, cells were re-suspended in ESC media lacking LIF and re-plated onto gelatin-coated plates or dishes. One micromole of *all-trans* retinoic acid was added to induce differentiation. Culture media were changed every day with fresh retinoic-acid-supplemented media without LIF. As controls, cells were cultured on gelatin-coated plates in the presence of LIF. For experiments in Fig. 3b–d, v6.5 cells were differentiated for 3 days.

To produce lentivirus, HEK293T cells were transfected with lentiviral vectors and packaging constructs (pLP1, pLP2 and pLP/VSV-G, Invitrogen) using Lipofectamine 2000. After 48 h, viral supernatant was collected and passed through 0.45-μm low-protein binding filters. To transduce v6.5, cells were plated on gelatin-coated plates in complete ESC media and incubated overnight at 37 °C, then viral supernatant was added together with 8 μg ml^{-1} polybrene. Six hours later, media was replaced with fresh ESC media. Transduction was repeated twice, followed by transferring cells to the plates containing mitomycin-C-treated puromycin-resistant mouse embryonic fibroblasts. Cells were selected on puromycin-resistant feeders with 2 μg ml^{-1} puromycin for at least 1 week.

Homology modelling of IDAX. Homology modelling of human IDAX protein was done on the Protein Model Portal server[37] (http://www.proteinmodelportal.org), which gives access to various comparative modelling methods provided by partner sites. In brief, IDAX (UniPort ID, Q9H2H0) was entered as a query, and a list of models generated from their respective template structures ranked by sequence homology was returned. Among the top two solutions, no. 1 is a ModBase model[38], whose template is the solution structure of the non-methyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase (PDB code, 2J2S). The modelled region shares 40% sequence identity when aligned with residues 127–179 of the target protein. No. 2 is generated by SWISS-MODEL[39] and is based on the crystal structure of the DNMT1–DNA complex (PDB code, 3PT6), which shares 35% sequence identity from residues 135–176 of the target protein. The alignment of the target sequence with its templates suggests that the TGHQ sequence in IDAX is part of a DNA-binding loop, because it is aligned with the RSKQ DNA-binding loop from DNMT1, and the IKKQ DNA-binding sequence in MLL. The superposition of the model with their respective templates was made using Coot[40], and the figure was created using Chimera[41].

31. Langmead, B. *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
32. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
33. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
34. Gardiner-Garden, M. *et al.* CpG islands in vertebrate genomes. *J. Mol. Biol.* **196,** 261–282 (1987).
35. Dreszer, T. R. *et al.* The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* **40,** D918–D923 (2012).
36. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
37. Arnold, K. *et al.* The protein model portal. *J. Struct. Funct. Genomics* **10,** 1–8 (2009).
38. Pieper, U. *et al.* MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **32,** D217–D222 (2004).
39. The SWISS-MODEL Repository. http://swissmodel.expasy.org/repository/.
40. Emsley, P. & Cowtan, K. *Coot*: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
41. Huang, C. C., Couch, G. S., Pettersen, E. F. & Ferrin, T. E. *Chimera: an Extensible Molecular Modeling Application Constructed Using Standard Components* Vol. 1, 724 (Pacific Symposium on Biocomputing, 1996).

LETTER

doi:10.1038/nature12121

Extensive transcriptional heterogeneity revealed by isoform profiling

Vicent Pelechano[1]*, Wu Wei[1,2]* & Lars M. Steinmetz[1,2]

Transcript function is determined by sequence elements arranged on an individual RNA molecule. Variation in transcripts can affect messenger RNA stability, localization and translation[1], or produce truncated proteins that differ in localization[2] or function[3]. Given the existence of overlapping, variable transcript isoforms, determining the functional impact of the transcriptome requires identification of full-length transcripts, rather than just the genomic regions that are transcribed[4,5]. Here, by jointly determining both transcript ends for millions of RNA molecules, we reveal an extensive layer of isoform diversity previously hidden among overlapping RNA molecules. Variation in transcript boundaries seems to be the rule rather than the exception, even within a single population of yeast cells. Over 26 major transcript isoforms per protein-coding gene were expressed in yeast. Hundreds of short coding RNAs and truncated versions of proteins are concomitantly encoded by alternative transcript isoforms, increasing protein diversity. In addition, approximately 70% of genes express alternative isoforms that vary in post-transcriptional regulatory elements, and tandem genes frequently produce overlapping or even bicistronic transcripts. This extensive transcript diversity is generated by a relatively simple eukaryotic genome with limited splicing, and within a genetically homogeneous population of cells. Our findings have implications for genome compaction, evolution and phenotypic diversity between single cells. These data also indicate that isoform diversity as well as RNA abundance should be considered when assessing the functional repertoire of genomes.

Transcript isoform variation and its functional relevance have been studied in detail for several single genes. For example, pluripotent cells express a dominant, truncated version of p53 that inhibits the function of the full protein, thereby promoting cell proliferation[3]. However, the genome-wide characterization of isoform variation has been limited. Identifying either 5′ or 3′ transcript boundaries individually[6–8] cannot determine the respective co-occurrence of start and end sites, which is essential for ascertaining the functional potential of a transcript. Thus, most studies have attributed variations at either transcript end to changes in the full-length messages. This interpretation is inaccurate in general, owing to transcripts that could arise from neighbouring genes, short abortive transcripts, bicistronic messages, and transcripts with differing lengths that overlap a gene. Thus, an important dimension of transcriptome complexity has remained largely unexplored. Here, we characterized the heterogeneity of transcript isoforms in *Saccharomyces cerevisiae* by jointly sequencing the 5′ and 3′ ends of each RNA molecule using an approach we term transcript isoform sequencing (TIF-Seq).

To capture *S. cerevisiae* transcript isoforms, capped and polyadenylated RNAs were converted into full-length complementary DNA molecules that were subjected to intramolecular ligation, fragmentation and capture of the 5′–3′ junctions through a biotin tag (Fig. 1a, Supplementary Fig. 1 and Supplementary Tables 1 and 2). The start and end sites of individual RNA molecules were then identified at single-nucleotide resolution by paired-end sequencing of the tagged fragments. We applied TIF-Seq to wild-type yeast grown in two conditions (with glucose (YPD) or galactose (YPGal) as the carbon source). We identified the exact 5′ cap and 3′ polyadenylation sites of more than 19 million individual RNA molecules (Fig. 1b, c). These transcripts are arranged in a remarkably complex, overlapping pattern across the genome (Fig. 1d). In addition to genes with variations in their untranslated regions (UTRs) (for example, *CBK1*), we discerned overlapping tandem genes (for example, *GIM3–YCK2*) and bicistronic transcripts (for example, *PGA1–IGO1*) (Fig. 1d). A comparison of our data with separate 5′ and 3′ end maps illustrates that the former cannot distinguish mono- from bicistronic transcripts, and the latter cannot distinguish 3′ UTR variation from short, overlapping 3′ end transcripts (for example, *YNL155W* and the antisense transcript of *YCK2* in Fig. 1d).

Altogether, in a genome containing only ~6,000 open reading frames (ORFs)[9], we detected over 1.88 million unique transcript isoforms (TIFs) (or 776,874 supported by at least two sequencing reads, Supplementary Data 1) that are defined by a unique combination of 5′ and 3′ end sites at single-nucleotide resolution. To enable analysis of major differences, we clustered the transcripts with each of their 5′ and 3′ end sites co-occurring within 5 nucleotides, and selected the highest expressed TIF per cluster as the representative mTIF (major transcript isoform, see Methods), yielding 371,087 mTIFs genome-wide (Supplementary Fig. 2 and Supplementary Data 2). This total corresponds to about half of all TIFs supported by two or more sequencing reads, demonstrating that there are both minor and substantial variations in transcript boundaries. Our further analysis uses TIFs and mTIFs supported by at least two sequencing reads. These numbers represent conservative estimates for transcript diversity, as our detection of isoforms is limited by sequencing depth, RNA abundance and length (Supplementary Information). We verified the accuracy of our transcript isoform mapping with extensive controls and independent confirmations (Supplementary Information, Supplementary Figs 3–8 and Supplementary Table 3).

Our data set reveals the extent to which different classes of transcripts are affected by isoform variation (Fig. 2a and Supplementary Fig. 9). We detected a median of 26 mTIFs (48 TIFs) that cover the coding region per verified or uncharacterized ORF in glucose and galactose (Fig. 2b–e). These mTIFs display a median positional variation of 75 nucleotides (26 for the 5′ start and 36 for the 3′ end, when considered independently) (Supplementary Fig. 10). Notably, this diversity is not dominated by a few highly abundant isoforms: a median of 10 mTIFs (or 29 TIFs) per gene is required to explain 80% of the mRNA population (Supplementary Fig. 11). Isoform heterogeneity is also found in non-coding genes, including an average of 7 mTIFs per stable unannotated transcript (SUT[4]) (Fig. 2b and Supplementary Fig. 12). In addition, we detected thousands of multicistronic mTIFs that cover two or more ORFs (Fig. 2a). Although the number of TIFs is probably higher than what we observed, we estimate a maximum of ~100 mTIFs (or 500 TIFs) per gene (Supplementary Fig. 13). Altogether,

[1]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]Stanford Genome Technology Center, Stanford University, Palo Alto, California 94304, USA.
*These authors contributed equally to this work.

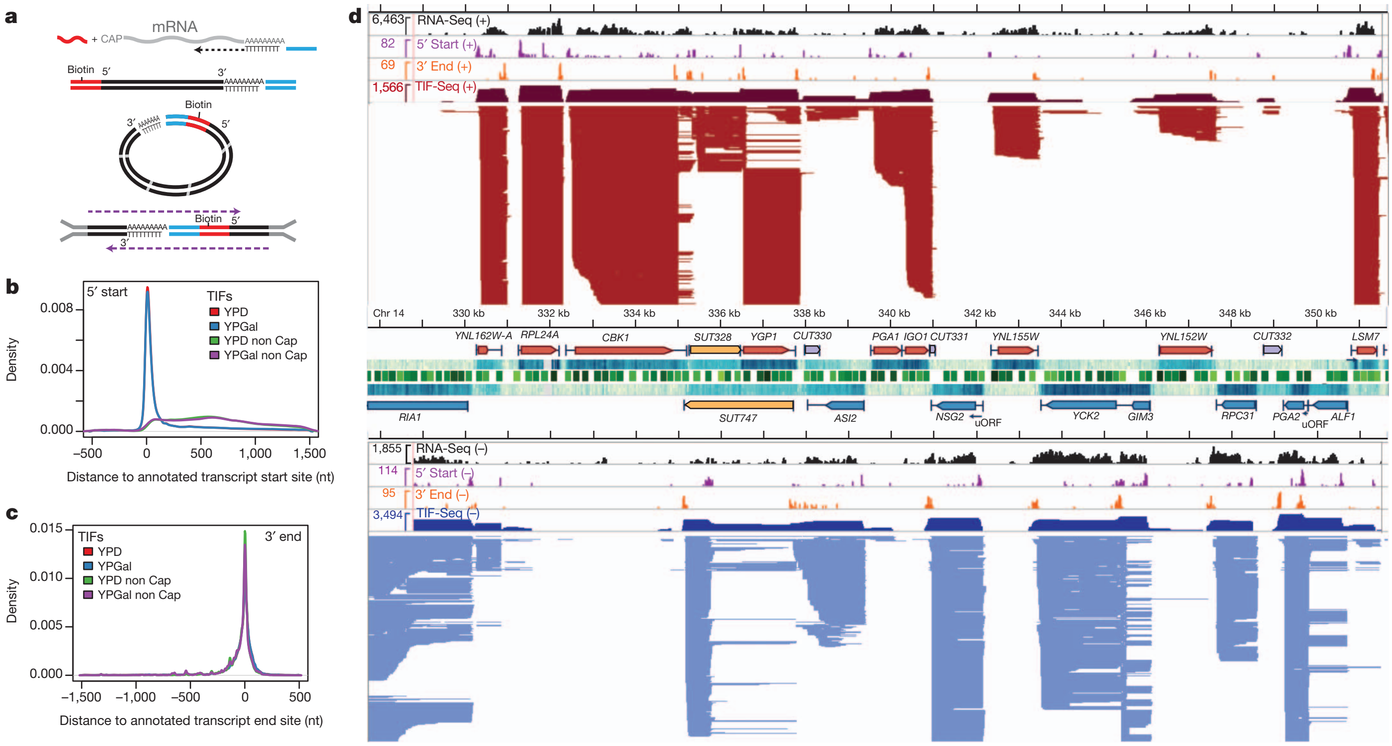


Figure 1 | Genome-wide measurement of transcript isoform diversity using TIF-Seq. **a**, The TIF-Seq method consists of RNA oligonucleotide capping, generation of full-length cDNA, circularization and paired-end sequencing. **b**, **c**, TIF boundaries agree overall with previous determinations of transcript 5′ starts (**b**) and 3′ ends (**c**) derived from tiling array annotations[4]. As expected, TIF-Seq of non-capped mRNAs does not produce many 5′ reads at the annotated transcript start sites (**b**). nt, nucleotides. **d**, Complex landscape of the yeast transcriptome in glucose, showing strand-specific RNA-Seq[25] in comparison to TIF-Seq 5′ start and 3′ end profiles, as well as TIF-Seq coverage in logarithmic scale (dark red/blue upper tracks); the maximum number of reads is indicated in each track. Individual TIFs are represented by red or blue lines (Watson (+) or Crick (−) strand, respectively), each line designating one TIF. Nucleosome positions (green track, darkness indicates significance[26]), expression measured by tiling arrays (blue heat map; darkness indicates expression level), and genome annotation[4] are shown in the centre: annotated ORFs (red and blue boxes for Watson and Crick strands, respectively), their UTRs (black lines), SUTs (yellow boxes), and CUTs (purple boxes). kb, kilobases. SUT, stable unannotated transcript; CUT, cryptic unstable transcript.

5,211 ORFs were covered by at least one mTIF, including 86% of verified or uncharacterized ORFs and 223 dubious ORFs[9] (Supplementary Data 3).

Most TIFs begin as expected: downstream of annotated transcription preinitiation complex (PIC) sites[10] and within the +1 nucleosome (Fig. 2c, Supplementary Discussion and Supplementary Figs 14–17). Notably, some interdependence between transcript start and end sites was observed in 382 protein-coding genes (false discovery rate (FDR) < 10%, Supplementary Fig. 18, Supplementary Data 4 and Supplementary Discussion), supporting the existence of interactions between promoters and terminators[11].

Although it is unclear how much of the isoform variation is functional, we discovered several cases where phenotypic consequences would be expected. First, we observed considerable variation in post-transcriptional regulatory elements. Most genes (66.9% in glucose, 71.9% across both conditions) with putative RNA-binding protein (RBP) sites[12] express mTIFs with different combinations of binding sites (Supplementary Fig. 19). Furthermore, RBP sites are enriched in regions that vary between isoforms ($P < 2.2 \times 10^{-16}$, Supplementary Information). Second, we observed significant variation in upstream ORFs (uORFs), short coding regions in the 5′ UTR that modulate translation efficiency of the downstream gene[13]. The standard understanding is that uORFs are transcribed along with the downstream gene, as both elements must be on the same RNA to interact. Yet over half of the genes with annotated uORFs[14] (703, 59% in Fig. 3a) expressed alternative mTIFs both with and without the uORF (for example, *ICY1* in Fig. 3a, b, Supplementary Data 5). This previously undetected occurrence, in addition to the variation in RNA binding sites, exemplifies transcriptional control of post-transcriptional regulatory potential: the precise isoform transcribed dictates the regulation that can be imposed on the gene.

Notably, our data set reveals that uORFs are not only translational regulators, but can also have an independent identity. Isoforms containing only the uORF were detected for 48% (567) of genes with known uORFs (Fig. 3c and Supplementary Fig. 20). Using ribosome profiling data[15], we found that genes containing genuine uORFs (where the mTIFs always span both the uORF and the main ORF) are significantly less translated than those where the uORFs are in fact independent, misannotated transcripts ($P < 2 \times 10^{-4}$) (Fig. 3d). This is consistent with the expected absence of translational repression by uORFs in the latter case. In addition to re-annotating uORFs, we detected the first downstream ORFs (dORFs), defined as short coding sequences within TIFs that also cover the upstream coding gene (for example, *COX19*, Supplementary Fig. 7 and Supplementary Data 6).

We confirmed the existence of several short transcripts previously misannotated as uORFs by northern blot (for example, *PCL7*, Fig. 3e and Supplementary Fig. 8). The fact that these transcripts have a canonical mRNA structure (5′ capped and polyadenylated), are bound by ribosomes[14], and are evolutionarily conserved (Supplementary Fig. 21) suggests that they are new short coding RNAs (scRNAs, Supplementary Data 5). Short peptides can perform crucial functions, as has recently been described in cellular differentiation[16]. The capacity of TIF-Seq to detect potentially peptide-encoding scRNAs opens new avenues for studying their function and regulation.

We also analysed the impact of transcript variation on protein diversity. Previous studies have identified alternative transcript isoforms that skip the first start codon, leading to loss of amino-terminal signal peptides and to alternative protein localization (for example,

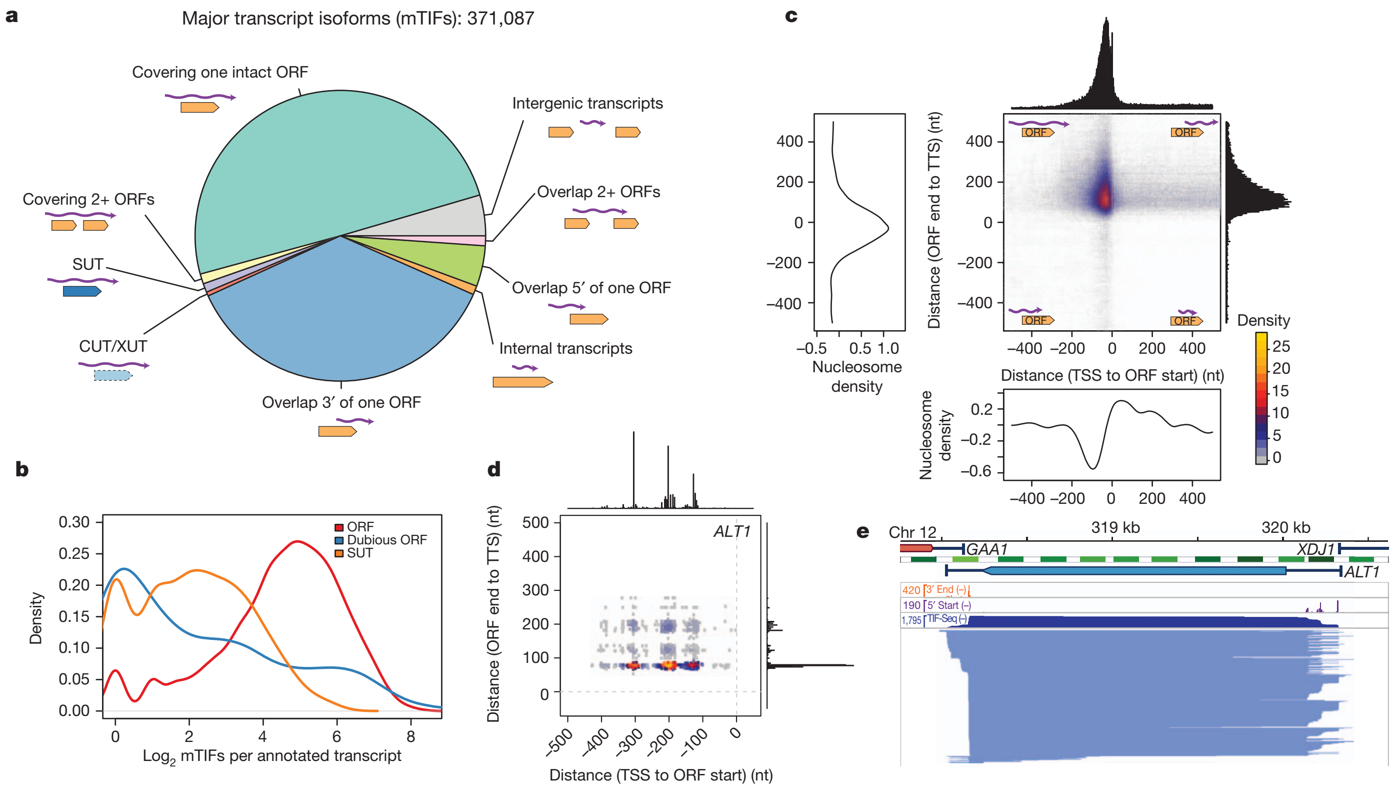


Figure 2 | Extensive isoform diversity revealed among overlapping RNA populations, both at the genomic and single-gene level. a, Categories of mTIFs identified in glucose and galactose. XUT, *XRN1*-sensitive unstable transcript. **b**, Log_2-scale distribution of clustered mTIFs per annotated transcript that cover characterized or uncharacterized ORFs (ORFs), dubious ORFs, or overlap more than 80% of stable unannotated transcripts (SUTs)[4]. **c**, Transcript end distance to ORF stop codon (*y* axis) versus transcript start distance to ORF start codon (*x* axis) genome-wide, revealing that most mTIFs cover the entire ORF. Decreased nucleosome density[27] coincides with peaks in transcript start and end site distributions. TTS, transcript termination site; TSS, transcript start site. **d**, Boundaries of TIFs covering *ALT1* relative to ORF boundaries (as in **c**). **e**, Structure of TIFs overlapping *ALT1* in glucose. 5′ start, 3′ end and TIF-Seq coverage in natural scale. Nucleosome and genome annotations as in Fig. 1d.

SUC2 (ref. 2), whose protein product can be either cytosolic or secreted, and *VAS1* (ref. 17)). We identified 153 additional genes in which at least half of the TIFs with coding potential skipped the first start codon (Fig. 4a and Supplementary Data 7). The translation of these truncated isoforms is supported by recent ribosome profiling data[15] (Fig. 4b and Supplementary Fig. 22), which along with recent proteomics data[18] indicate that N-terminal truncation via alternative start codon usage is a common phenomenon. This phenomenon can be transcriptionally regulated: we detected 9 genes with significantly differential truncation between glucose and galactose ($P < 10^{-3}$, FDR < 0.1, Fig. 4c and Supplementary Table 4). These findings indicate a more common production of truncated 5′ transcripts than previously appreciated, which can lead directly to increased protein diversity even without post-transcriptional regulation.

Our data set also provides evidence for the production of carboxyl-terminal protein truncation via alternative polyadenylation sites. We identified 33 genes enriched for internal polyadenylation that introduces early stop codons into the RNA that are not encoded by the DNA[19] (FDR $< 10\%$, Supplementary Discussion, Supplementary Fig. 23, Supplementary Table 5 and Supplementary Data 8). Among them is *GAL10* ($P < 1.2 \times 10^{-9}$, Fig. 4d), which encodes a bifunctional enzyme in *S. cerevisiae* with two enzymatic domains that are encoded by two separate genes in other organisms[20]. In galactose media, such early stop codons result in additional transcripts encoding proteins with only one of these domains (Fig. 4d). Our evidence of protein truncation via alternative isoform usage represents a plausible means for organisms such as yeast, in which alternative splicing is uncommon, to increase protein diversity by selective domain truncation.

Our genome-wide map of transcript boundaries enabled us to measure the extent of transcriptional compaction on each strand of the genome. Most tandem TIFs are separated by approximately 150 base pairs (bp; Supplementary Fig. 24). However, chained arrangements between adjacent TIFs are common, where the end of one TIF coincides with the start of the TIF for the downstream gene. In fact, of 2,747 tandem ORF pairs in the genome, 27% (743) express overlapping mTIFs (for example, *GIM3–YCK2*, Fig. 1d and Supplementary Data 9) and 6.7% (185) produce bicistronic transcripts (Supplementary Data 10). Most overlapping transcripts stop within the first 100–200 bp of the downstream gene (Supplementary Discussion and Supplementary Fig. 24), indicating that upstream elongating and downstream initiating RNA polymerases are distinguished within this window[21]. This common overlap facilitates crosstalk between transcriptional units, wherein the expression of a gene can depend not only on its own promoter, but also on the expression of its neighbours[22].

Our data set reveals the extent of transcript isoform diversity in the yeast genome at unprecedented resolution. Since most yeast genes have fewer than one mRNA molecule per cell[23], the sheer number of isoforms detected here, even within a single environmental condition, indicates that every cell in a clonal population has a unique transcriptome in terms of RNA abundance, sequence and thus regulatory potential. Such cell-to-cell heterogeneity may confer evolutionary advantages, enabling more rapid adaptation of the species to unforeseen environmental challenges. The variation in transcript isoforms has functional consequences through its impact on post-transcriptional regulatory potential, as well as on protein length and localization. In addition, we discovered hundreds of short coding RNAs whose function can now be investigated. Further applications of the TIF-Seq method, or of alternative paired-end strategies such as RNA-PET[24], to additional environmental conditions, genetic backgrounds and organisms will deepen our understanding of transcriptional complexity.

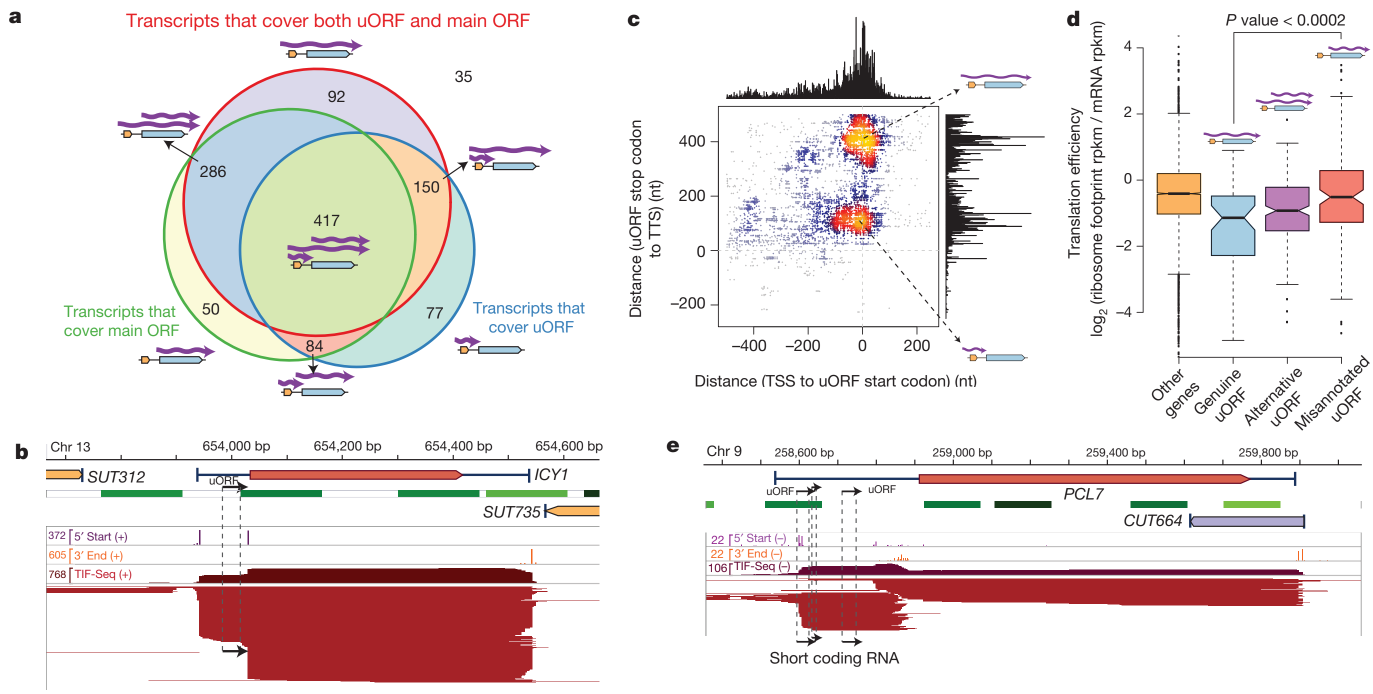


Figure 3 | Transcript isoforms with varying regulatory elements and independent short coding RNAs. **a**, Number of genes whose mTIFs overlap with previously annotated upstream ORFs (uORFs) and their associated (main) ORFs[14]. **b**, *ICY1* transcripts in glucose show alternative presence of uORFs (marked with arrows). **c**, Genome-wide plot of uORF-containing mTIFs: transcript end distance to uORF stop codon (*y* axis) versus transcript start distance to uORF start codon (*x* axis). Small coding RNAs previously misannotated as uORFs represent a separate population of short overlapping RNAs. **d**, Genes with mTIFs that always contain uORFs have lower translation efficiency[15] than those for which the uORF is independently transcribed. Genes with alternative presence of uORFs (for example, *ICY1*) have intermediate translation efficiency. Significance was computed using the Wilcoxon rank-sum test with continuity correction. **e**, Example of an scRNA that was previously misannotated as a uORF in the *PCL7* locus (glucose data shown).

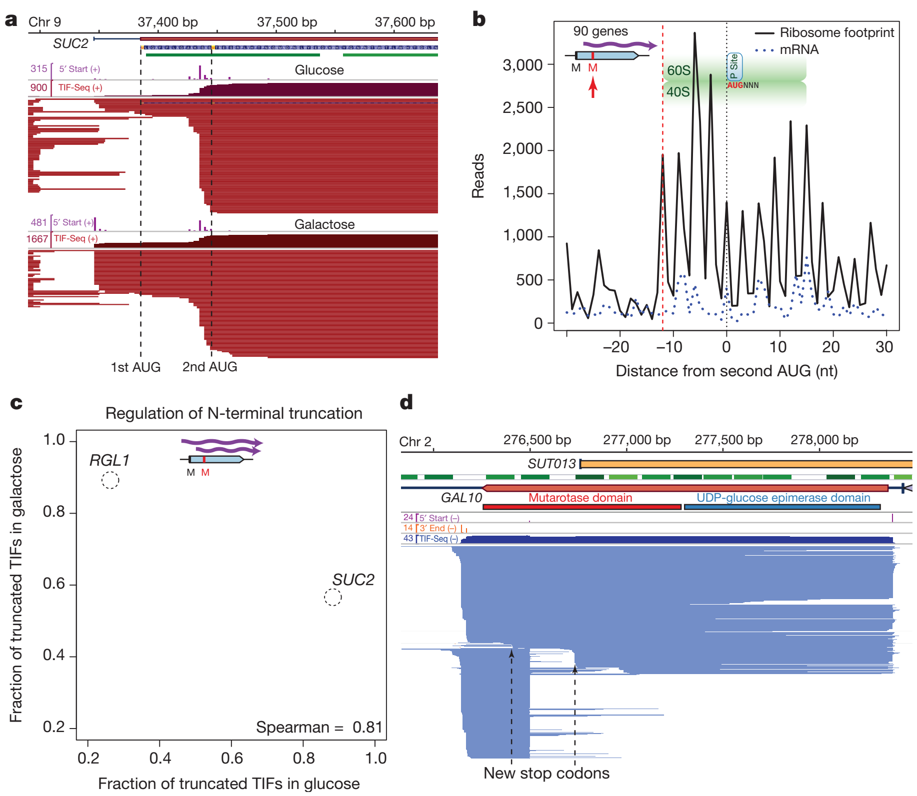


Figure 4 | Alternative transcript isoforms increase coding diversity. **a**, Differential isoform regulation between glucose and galactose produces alternative truncated proteins with differential cellular localization, as shown here for *SUC2* (ref. 2). This regulation is due to subtle variations in TSS selection (5′ start track in purple) that result in alternative inclusion of the first AUG. **b**, Genes producing truncated transcripts that skip the first AUG (80% of these TIFs start between the first and second AUG) are effectively translated and show the expected codon usage pattern and ribosomal protection (green) in ribosome profiling data[15], starting at but not before the second in-frame methionine codon. **c**, Proportion of N-terminal truncated TIFs, (that is, using the second methionine as start codon) in glucose and galactose. **d**, TIFs with internal polyadenylation events that introduce novel stop codons into the RNA encode truncated ORFs and potentially alternative protein isoforms, as shown here for *GAL10*.

In multicellular organisms, the combination of transcript boundary variation and alternative splicing is expected to amplify the diversity of transcript isoforms generated from a single genomic sequence, thereby expanding the functional repertoire of the genome.

METHODS SUMMARY

TIF-Seq library construction was performed as described in Methods and Supplementary Methods. Libraries (Supplementary Table 1) were sequenced using an Illumina HiSeq 2000. Sequencing read analysis was performed using R and Bioconductor (http://www.bioconductor.org/). Only TIFs and mTIFs supported by at least two sequencing reads were used for statistical analysis. Genome sequences (SGD R64) and annotation feature files for S288c were obtained from SGD on 26 March 2011 (http://www.yeastgenome.org/).

Full Methods and any associated references are available in the online version of the paper.

Received 18 December 2012; accepted 26 March 2013.
Published online 24 April 2013.

1. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* **43,** 853–866 (2011).
2. Carlson, M. & Botstein, D. Two differentially regulated mRNAs with different 5′ ends encode secreted with intracellular forms of yeast invertase. *Cell* **28,** 145–154 (1982).
3. Ungewitter, E. & Scrable, H. Δ40p53 controls the switch from pluripotency to differentiation by regulating IGF signaling in ESCs. *Genes Dev.* **24,** 2408–2419 (2010).
4. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457,** 1033–1037 (2009).
5. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
6. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
7. Ozsolak, F. *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143,** 1018–1029 (2010).
8. Zhang, Z. & Dietrich, F. S. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. *Nucleic Acids Res.* **33,** 2838–2851 (2005).
9. Cherry, J. M. *et al. Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40,** D700–D705 (2012).
10. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483,** 295–301 (2012).
11. Tan-Wong, S. M. *et al.* Gene loops enhance transcriptional directionality. *Science* **338,** 671–675 (2012).
12. Riordan, D. P., Herschlag, D. & Brown, P. O. Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome. *Nucleic Acids Res.* **39,** 1501–1509 (2011).
13. Hood, H. M., Neafsey, D. E., Galagan, J. & Sachs, M. S. Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi. *Annu. Rev. Microbiol.* **63,** 385–409 (2009).
14. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
15. Gerashchenko, M. V., Lobanov, A. V. & Gladyshev, V. N. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl Acad. Sci. USA* **109,** 17394–17399 (2012).
16. Kondo, T. *et al.* Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* **329,** 336–339 (2010).
17. Chatton, B., Walter, P., Ebel, J. P., Lacroute, F. & Fasiolo, F. The yeast VAS1 gene encodes both mitochondrial and cytoplasmic valyl-tRNA synthetases. *J. Biol. Chem.* **263,** 52–57 (1988).
18. Fournier, C. T. *et al.* Amino termini of many yeast proteins map to downstream start codons. *J. Proteome Res.* **11,** 5712–5719 (2012).
19. Yao, P. *et al.* Coding region polyadenylation generates a truncated tRNA synthetase that counters translation repression. *Cell* **149,** 88–100 (2012).
20. Majumdar, S., Ghatak, J., Mukherji, S., Bhattacharjee, H. & Bhaduri, A. UDPgalactose 4-epimerase from *Saccharomyces cerevisiae*. A bifunctional enzyme with aldose 1-epimerase activity. *Eur. J. Biochem.* **271,** 753–759 (2004).
21. Mayer, A. *et al.* CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* **336,** 1723–1725 (2012).
22. Xu, Z. *et al.* Antisense expression increases gene expression variability and locus interdependency. *Mol. Syst. Biol.* **7,** 468 (2011).
23. Miura, F. *et al.* Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC Genomics* **9,** 574 (2008).
24. Ruan, X. & Ruan, Y. Genome wide full-length transcript analysis using 5′ and 3′ paired-end-tag next generation sequencing (RNA-PET). *Methods Mol. Biol.* **809,** 535–562 (2012).
25. Wilkening, S. *et al.* An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res.* **41,** e65 (2013).
26. Venters, B. J. & Pugh, B. F. A canonical promoter organization of the transcription machinery and its regulators in the *Saccharomyces* genome. *Genome Res.* **19,** 360–371 (2009).
27. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–366 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Aiyar for help in editing and refining the manuscript. We thank W. Huber, C. Zhu, A. I. Järvelin, S. Clauder-Münster, J. Zaugg, S. Adjalley, G. Lin and the members of the Steinmetz laboratory for helpful discussions and critical comments on the manuscript. We thank V. N. Gladyshev and C. Pineau for sharing published data. This study was technically supported by the EMBL Genomics Core Facility. This study was financially supported by the National Institutes of Health (to L.M.S.). V.P. was supported by an EMBO fellowship.

Author Contributions W.W., V.P. and L.M.S. conceived the project. V.P. developed the TIF-Seq method and performed experiments. W.W. and V.P. performed the analysis. V.P., W.W. and L.M.S. wrote the manuscript.

Author Information The data reported in this paper have been deposited in GEO under accession number GSE39128 and are also accessible at http://steinmetzlab.embl.de/TIFSeq. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.M.S. (larsms@embl.de).

METHODS

Biological samples. *S. cerevisiae* strain SLS045 (*MAT***a**/α *GAL2/GAL2*, S288c background) was grown to mid-log phase ($D_{600\,nm} \approx 1$) using either YPD (1% yeast extract, 2% peptone, 2% glucose) or YPGal (1% yeast extract, 2% peptone, 2% galactose). Total RNA was isolated by the standard hot phenol method and contaminant DNA was removed by DNase I treatment. Sequenced samples are shown in Supplementary Table 1.

TIF-Seq method. For the construction of the TIF-Seq libraries, we used 60 μg of DNA-free total RNA as input. As an internal control, we added capped and polyadenylated *in vitro* transcripts (ATCC 87482, 87483 and 87484). The 5′ end of the non-capped RNA molecules was dephosphorylated by treatment with 6 units of shrimp alkaline phosphatase (SAP, Fermentas) for 30 min at 37 °C in the presence of RNase inhibitor (RNasin +, Promega). The RNA was then purified by a double phenol extraction and ethanol precipitation. CAP was removed by a 1-h incubation at 37 °C with 5 units of tobacco acid pyrophosphatase (TAP, Epicentre) in the presence of RNase inhibitor. The sample was phenol:chloroform-purified and ethanol-precipitated. Finally, for oligonucleotide ligation to the 5′ end of the formerly capped molecules, the treated RNA sample was incubated overnight at 16 °C with 20 units of T4 RNA ligase 1 (NEB) in the presence of 10 mM DNA/RNA '5oligocap' oligonucleotide (Supplementary Table 2), 10% dimethylsulphoxide (DMSO) and RNase inhibitor. The RNA was column-purified (RNeasy, Qiagen) and its integrity was checked (Bioanalyzer, Agilent).

To control for the presence of chimaeras, each ligated RNA sample was divided into two aliquots and independently processed to generate two fractions of full-length cDNA (FlcDNA) with different terminal barcoding (see Supplementary information for details). Specifically, each fraction (11.2 μl) was mixed with 1 μl RT priming oligonucleotide, 1 μM either 3cDNANotI_A or 3cDNANotI_B (Supplementary Table 2), and 1 μl 10mM dNTPs. The sample was incubated at 65 °C for 5 min and transferred to ice. 4 μl of 5× First-Strand buffer (Invitrogen), 2 μl DTT 0.1 M and 0.5 μl RNase inhibitor were added to each sample, which was incubated at 42 °C for 2 min to minimize possible mispriming. 2 μl of Superscript III reverse transcriptase (200 U μl^{-1}, Invitrogen) and high temperature (55 °C) incubation was then used for the retrotranscription to minimize the effects of RNA secondary structure. The reaction was incubated for 20 min at 42 °C, 40 min at 50 °C, and 20 min at 55 °C with a final 15 min enzyme inactivation at 70 °C. RNA removal was performed by adding 0.5 μl RNase cocktail (Ambion) and 2.5 units of RNase H (NEB) to each sample and incubating at 37 °C for 30 min. Samples were purified using Agencourt AMPure XP beads (Beckman Coulter Genomics) according to the manufacturer's instructions and eluted in 19 μl EB (10 mM TrisHCl pH 8.0). 19 μl of resulting FlcDNA samples were PCR-amplified using 20 μl 2× HF Phusion MasterMix (Finnzymes), 0.5 μl biotinylated oligonucleotide 5 μM (either 5BioNotI_A or 5BioNotI_B) and 0.5 μl oligonucleotide 5 μM (either 3AmpNotI_A or 3AmpNotI_B) (Supplementary Table 2). The following thermocycler program was used: 30 s for initial denaturation at 98 °C, 10 cycles (20 s of denaturation at 98 °C, 30 s of annealing at 50 °C (+1 °C per cycle) and 5 min elongation at 72 °C (+10 s per cycle)) and a final elongation of 5 min at 72 °C. Samples were purified using AMPure XP beads. The two independently barcoded aliquots were ultimately pooled together.

To generate cohesive ends, FlcDNA samples were digested for 1 h at 37 °C with 100 units of NotI (NEB) and heat-inactivated for 20 min at 65 °C. The samples were AMPure XP-purified and DNA yield was quantified. Between 300 and 600 ng of digested FlcDNA was circularized for 16 h at 16 °C by intramolecular ligation with 20 μl of T4 DNA ligase (2,000 units μl^{-1}, NEB) in 600 μl final volume. Non-circularized molecules were degraded by incubating the samples for 20 min at 37 °C with 20 units each of exonuclease III (NEB) and exonuclease I (NEB). Enzymes were inactivated by adding 12 μl of 0.5 M EDTA and incubating the samples at 70 °C for 30 min. Circularized FlcDNAs were then phenol:chloroform-purified and ethanol-precipitated.

Purified, circularized FlcDNAs were resuspended in 130 μl EB and sonicated with a Covaris S220 (4 min, 20% duty cycle, intensity 5, 200 cycles per burst). The fragmented DNA was purified with AMPure XP beads and eluted with 20 μl EB. Biotin-containing fragments were captured by incubating the samples for 30 min at room temperature with 20 μl of Streptavidin-conjugated Dynabeads M-280 (Invitrogen) and washed according to the manufacturer's instructions.

Addition of forked barcoded adapters to the captured molecules was performed using the standard Illumina DNA-Seq library generation protocol with some minor modifications. Specifically, purifications using AMPure beads were replaced with separation on magnetic Dynabeads and NEBNext Master Mixes (NEB) were used. A 20-cycle PCR enrichment was performed using Phusion polymerase (Finnzymes). 300-bp libraries were isolated using e-Gel 2% SizeSelect (Invitrogen) and sequenced with a HiSeq 2000 (Illumina) using paired-end sequencing of 105 bp reads.

TIF-Seq method for long mRNA molecules. We used the same method as described above, but introduced an additional size selection step. Specifically, after the initial PCR amplification, the FlcDNA samples were size-selected on a 1.5% agarose gel, and fragments over 2 kb were purified using QIAquick Gel Extraction Kit (Qiagen). The recovered FlcDNA samples were reamplified using 10 cycles of PCR before the NotI digestion.

TIF-Seq method for non-capped mRNA molecules (mono- and triphosphorylated mRNAs). We used the same method as described above, but modified steps before the ssRNA ligation. RNA was dephosphorylated using shrimp alkaline phosphatase as described above, but instead of proceeding to treatment with tobacco acid pyrophosphatase, RNA was rephosphorylated for 1 h at 37 °C using T4 polynucleotide kinase (NEB).

RNA circularization and targeted sequencing. Sixty micrograms of DNA-free RNA samples were SAP-and TAP-treated as described above to obtain full-length RNA molecules with 5′ phosphate ends. RNA circularization was performed in the presence of RNase inhibitor, 10% DMSO, T4 RNA ligase buffer, and 50 units of T4 RNA ligase 1 (ssRNA ligase, NEB) for 16 h at 16 °C. The circularized RNA was purified with RNeasy columns (Qiagen) and subjected to random hexamer retrotranscription. The resulting cDNAs were used as a template for standard PCR amplification with divergent oligonucleotides (Supplementary Table 2). The PCR products were cloned using the TOPO TA cloning system (Invitrogen). Individual clones were bidirectionally sequenced with Sanger sequencing to determine both 5′ and 3′ ends. Only clone sequences spanning a poly(A)-tail were taken into consideration.

Northern blot. The DIG Starter Northern kit (Roche) was used according to the manufacturer's instructions. Strand-specific RNA-DIG probes were generated by *in vitro* transcription.

Sequencing read processing and alignment. Sequencing reads were de-multiplexed and barcode sequences were removed. The presence of internal chimaera control barcodes was assessed by the Needleman–Wunsch global alignment method provided by the R Biostrings package from Bioconductor (http://www.bioconductor.org/). Samples were classified into 4 groups (putative intermolecular (A-A and B-B) or intramolecular events (A-B and B-A)). Only high-confidence reads with both chimaera control barcodes and a poly(A)-tail were considered for further analysis.

Pairs of 5′ sequences and 3′ sequences were trimmed and then separately aligned to the reference genome using Novoalign V2.07.10 (http://www.novocraft.com) using default parameters. The S288c *S. cerevisiae* genome (SGD R64, http://www.yeastgenome.org), along with the sequences of the *in vitro* transcripts that were included as spike-in controls, were used as reference sequences. Only sequences where both ends mapped to the reference were further analysed. Intermolecular pairs (A-B and B-A) were discarded. To exclude any other possible intermolecular cDNA species, only TIFs with both ends mapping to the same chromosome with a length ranging from 40 to 5,000 bp were considered for further analysis.

TIF clustering and mTIF definition. We clustered the transcripts with 5′ and 3′ end sites co-occurring within 5 bp (Supplementary Fig. 2a). Specifically, we defined TSS and TTS clusters separately. Each cluster was defined by both a window and a mTSS/TTS (the most abundant within that window). Clusters of TSS/TTSs were assigned iteratively in decreasing order of expression. In this process, each TSS/TTS site was compared to previously defined clusters, and the site was: (1) defined as a new mTSS/TTS with a 5-bp window (±2 bp up/downstream) if the window did not overlap with previous clusters; (2) defined as a new mTSS/TTS with a smaller window (<5 bp) to avoid overlap with previously defined clusters, if the TSS/TTS was ≥5 bp away from the closest previously defined mTSS/TTS but ≤2 bp from the closest cluster; and (3) merged with a previously defined cluster if the TSS/TTS was <5 bp away from the closest mTSS/TTS, in which case the original cluster window was extended to include the newly assigned TSS/TTS (thus the maximum window size of one cluster would be 9 bp); if the TSS/TTS overlapped with 2 previously defined mTSS/TTS in <5 bp, it was merged with the cluster with the closer (or higher expressed) mTSS/TTS. After this assignment process, only clusters defined by mTSS/TTSs with at least 3 supporting reads were considered. mTIFs were defined as connections between mTSS and mTTS supported by at least 2 reads connecting the associated clusters. All TIFs that shared a given TSS/TTS cluster were assigned to the corresponding mTIF cluster.

TIF annotation. The TIFs were aligned to genome annotation features and classified as: (1) ORFs, if they covered an intact ORF coding region; (2) bicistronic TIFs, if they covered 2 or more ORFs; (3) SUTs, if the common region between the TIF and SUT included more than 80% of the length of both the TIF and the SUT; (4) CUTs/XUTs, same as SUTs; (5) overlapping 2 ORFs, if they overlapped two ORFs but did not entirely cover either; (6) overlapping 5′ of one ORF, if they overlapped only the 5′ of one ORF; (7) overlapping 3′ of one ORF, if they overlapped only the 3′ of one ORF; and (8) intergenic TIF, if

they did not overlap with any annotated ORFs or overlapped with less than 80% of annotated SUTs, CUTs or XUTs. Annotated transcripts (ORF-Ts, SUTs and CUTs) with TSSs and TTSs are from our previous study that used tiling arrays[4].

Comparing TSSs and TTSs to nucleosome data. Nucleosome raw data are derived from ref. 27. Normalized nucleosome occupancy values for the regions flanking the TSSs or TTSs (±500 bp) were extracted and the median values in each position were calculated and plotted. TSSs or TTSs from mTIFs were used.

doi:10.1038/nature12061

The catalytic mechanism for aerobic formation of methane by bacteria

Siddhesh S. Kamat[1], Howard J. Williams[1], Lawrence J. Dangott[2], Mrinmoy Chakrabarti[1] & Frank M. Raushel[1]

Methane is a potent greenhouse gas that is produced in significant quantities by aerobic marine organisms[1]. These bacteria apparently catalyse the formation of methane through the cleavage of the highly unreactive carbon–phosphorus bond in methyl phosphonate (MPn), but the biological or terrestrial source of this compound is unclear[2]. However, the ocean-dwelling bacterium *Nitrosopumilus maritimus* catalyses the biosynthesis of MPn from 2-hydroxyethyl phosphonate[3] and the bacterial C–P lyase complex is known to convert MPn to methane[4–7]. In addition to MPn, the bacterial C–P lyase complex catalyses C–P bond cleavage of many alkyl phosphonates when the environmental concentration of phosphate is low[4–7]. PhnJ from the C–P lyase complex catalyses an unprecedented C–P bond cleavage reaction of ribose-1-phosphonate-5-phosphate to methane and ribose-1,2-cyclic-phosphate-5-phosphate. This reaction requires a redox-active [4Fe–4S]-cluster and *S*-adenosyl-L-methionine, which is reductively cleaved to L-methionine and 5′-deoxyadenosine[8]. Here we show that PhnJ is a novel radical *S*-adenosyl-L-methionine enzyme that catalyses C–P bond cleavage through the initial formation of a 5′-deoxyadenosyl radical and two protein-based radicals localized at Gly 32 and Cys 272. During this transformation, the *pro-R* hydrogen from Gly 32 is transferred to the 5′-deoxyadenosyl radical to form 5′-deoxyadenosine and the *pro-S* hydrogen is transferred to the radical intermediate that ultimately generates methane. A comprehensive reaction mechanism is proposed for cleavage of the C–P bond by the C–P lyase complex that uses a covalent thiophosphate intermediate for methane and phosphate formation.

The glutathione *S*-transferase (GST) fusion protein of PhnJ from *Escherichia coli* was purified under anaerobic conditions[8]. The isolated protein was dark brown in colour, had an absorbance maximum at a wavelength of 410 nm and was EPR silent (produced no electron paramagnetic resonance signal) (Fig. 1a). Addition of dithionite to the isolated protein resulted in the loss of absorbance at 410 nm and yielded an EPR-active species (Fig. 1a). At a temperature of 12 K the EPR signal was strongest, and at 48 K the signal was significantly weaker (Supplementary Fig. 1). These results are consistent with the initial isolation of an intact $[4Fe–4S]^{2+}$-cluster that can be further reduced by dithionite to the $[4Fe–4S]^{1+}$ oxidation state[9–12].

Iron–sulphur cluster formation in most radical *S*-adenosyl-L-methionine (SAM) enzymes requires coordination to three cysteine residues in a CX_3CX_2C motif[13–15] (X, any amino acid). PhnJ lacks the signature radical SAM enzyme motif but has four cysteine residues with a $CX_2CX_{21}CX_5C$ spacing near the carboxy-terminal end of the protein. To determine which three of the four cysteine residues of PhnJ are required for assembly of the [4Fe–4S]-cluster, we mutated Cys 241, Cys 244, Cys 266 and Cys 272 to Ala. All of the mutant enzymes were inactive for the production of methane, and the Cys241Ala, Cys244Ala and Cys266Ala mutants were unable to assemble a [4Fe–4S]-cluster (Supplementary Fig. 2). The Cys272Ala mutant protein was dark brown in colour and the ultraviolet–visible spectrum was identical to that of wild-type PhnJ (Fig. 1b). After reduction of the Cys272Ala mutant protein with dithionite, the absorbance maximum at 410 nm was lost and an EPR-active species was formed that is identical to that of wild-type PhnJ (Fig. 1b). Cys 241, Cys 244 and Cys 266 are therefore required for the formation of the [4Fe–4S]-cluster in PhnJ, and Cys 272 is critical for catalytic activity.

It was shown previously that 5′-deoxyadenosine (Ado-CH_3) and L-methionine are formed from the utilization of SAM during the reaction catalysed by PhnJ and that approximately one enzyme equivalent of SAM is consumed under single or multiple turnovers[8]. The reductive cleavage of SAM by the $[4Fe–4S]^{1+}$-cluster requires the transient formation of a 5′-deoxyadenosyl radical (Ado-CH_2•) that subsequently abstracts a hydrogen atom from either the enzyme or the substrate as

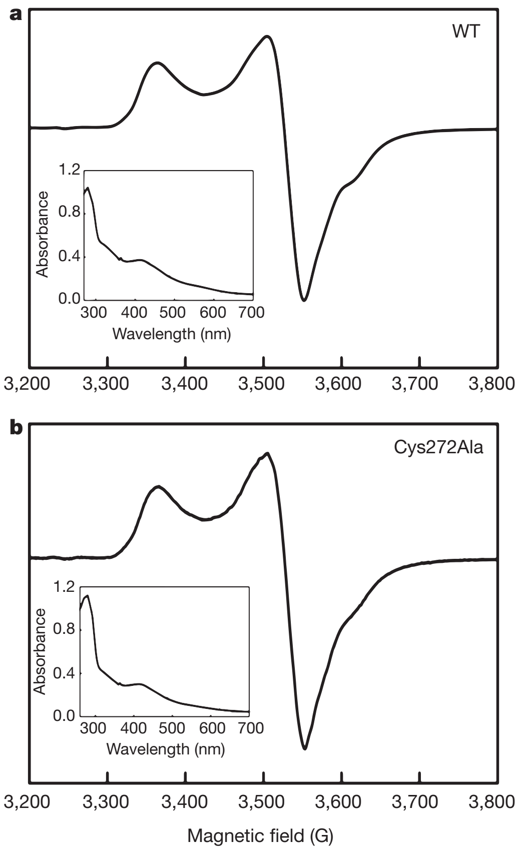
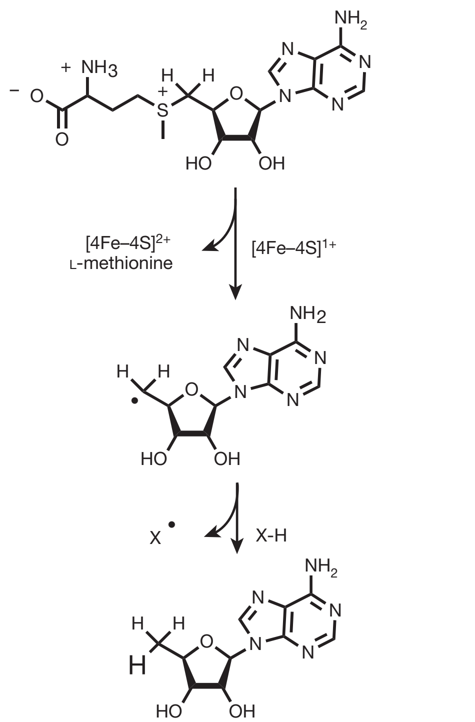


Figure 1 | EPR spectra of wild-type PhnJ and the Cys272Ala mutant. a, EPR spectrum of wild-type (WT) PhnJ (180 μM) after reduction with dithionite at 12 K. The EPR spectrum is characteristic of a reduced $[4Fe–4S]^{1+}$-cluster with *g* values of 2.01, 1.92 and 1.87. Inset, ultraviolet–visible spectrum of as-isolated wild-type PhnJ (18 μM). **b**, EPR spectrum of PhnJ Cys272Ala mutant (158 μM) after reduction with dithionite at 12 K. Inset, ultraviolet–visible spectrum of as-isolated PhnJ Cys272Ala mutant (17 μM). The EPR spectra were obtained under these instrument settings: 9.46-GHz microwave frequency, 0.2-mW microwave power, and 10-G modulation amplitude.

[1]Department of Chemistry, PO Box 30012, Texas A&M University, College Station, Texas 77843, USA. [2]Protein Chemistry Laboratory, Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas 77843, USA.

illustrated in equation (1)[13–15] (where X-H denotes the enzyme or substrate and H denotes the abstracted hydrogen atom). To determine whether or not hydrogen atom abstraction occurs from a solvent exchangeable site, the reaction catalysed by PhnJ was conducted in H_2O and then in 90% D_2O. The product, Ado-CH_3, was isolated and the isotopic composition determined by mass spectrometry. For the reaction conducted in H_2O, the mass of the $(M+H)^+$ ion of the isolated Ado-CH_3 was 252.1 Da. (Fig. 2a). When the reaction was performed in 90% D_2O, the $(M+H)^+$ ion mass was also 252.1 Da (Fig. 2b). These results demonstrate that the hydrogen atom that is transferred to the Ado-CH_2• radical does not originate in a solvent exchangeable site in either the protein or the substrate.



(1)

Because hydrogen atom transfer to the Ado-CH_2• radical intermediate does not occur from a solvent exchangeable site, the next most probable source of this hydrogen atom was postulated to be a Gly residue[16–19]. PhnJ was therefore expressed and purified from an M9-minimal medium supplemented with [2,2-$^2H_2$]-Gly. The mass spectrum (Supplementary Fig. 3) of a typical tryptic peptide (^{164}FGHIATTY AYPVK^{176}) demonstrated that the average deuterium content of the Gly residues was as follows: PhnJ-Gly-h_2, 19%; PhnJ-Gly-hd, 15%; PhnJ-Gly-d_2, 66%. When the Gly-labelled protein was used to catalyse the C–P lyase reaction in H_2O, the Ado-CH_3 had a predominant $(M+H)^+$ ion mass of 253.1 Da and the deuterium content of the newly formed Ado-CH_3 was ~74% (Fig. 2c). Therefore, hydrogen atom transfer within PhnJ must occur from one of the eight conserved Gly residues to the transient Ado-CH_2• radical intermediate and consequently forms a glycyl radical.

During cleavage of the C–P bond of ribose-1-phosphonate-5-phosphate (PRPn) to ultimately form ribose-1,2-cyclic-phosphate-5-phosphate (PRcP), a new C–H bond is formed in the methane product, but the origin of the hydrogen atom is unknown (equation (2)). To determine the direct source of this hydrogen atom, the reaction catalysed by PhnJ was conducted in H_2O and D_2O under conditions of single and multiple turnovers of substrate using wild-type PhnJ, and PhnJ that was uniformly labelled with deuterated Gly. The methane produced in these reactions was trapped and subjected to mass spectrometric analysis to determine the ratio of unlabelled (CH_4) and deuterated (CH_3D) methane. When unlabelled PhnJ was incubated with less than one enzyme equivalent of substrate and the reaction

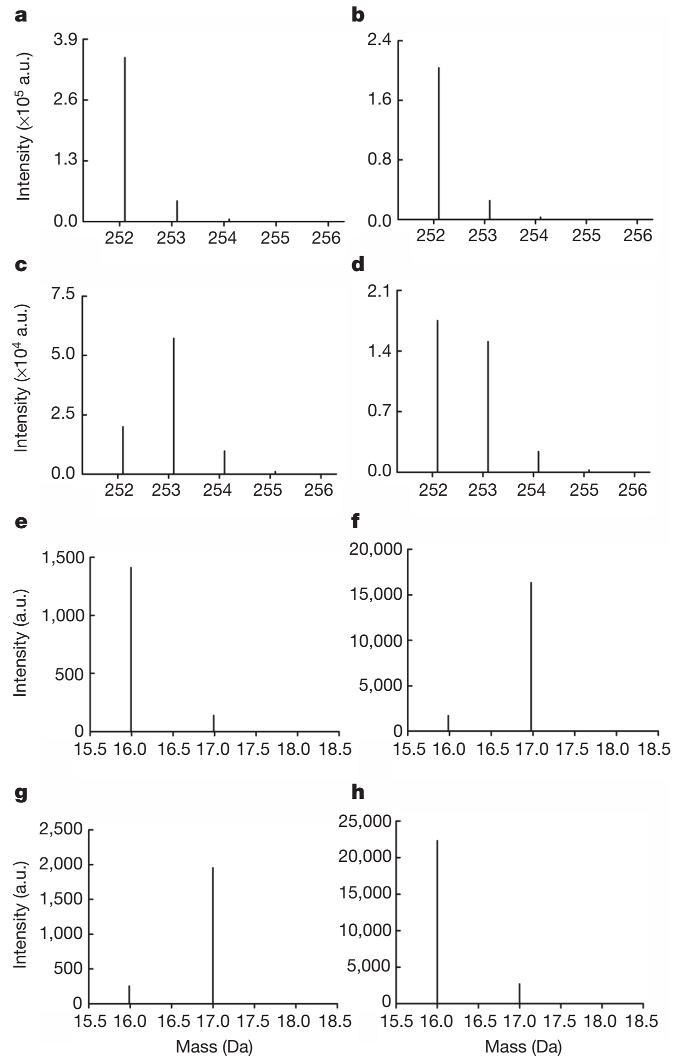
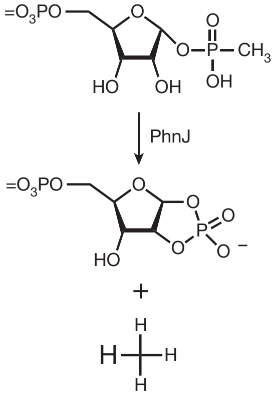


Figure 2 | Mass spectra of 5′-deoxyadenosine and methane from reactions catalysed by PhnJ. a–d, Mass spectra of 5′-deoxyadenosine for PhnJ in H_2O (**a**), PhnJ in 90% D_2O (**b**), PhnJ-Gly-d_2 in H_2O (**c**) and PhnJ-Gly-d_R in H_2O (**d**).Typical reaction compositions were 120 μM PhnJ, 1 mM PRPn (60 μM in **d**), 2 mM SAM, 1 mM dithionite, 50 units Factor Xa, 150 mM HEPES (pH 8.5), 0.5 M NaCl and 10% (w/v) glycerol. Typical reaction volume was 200 μl. **e–h**, Mass spectra of methane for wild-type PhnJ in 90% D_2O for a single-turnover experiment (**e**), wild-type PhnJ in 90% D_2O for multiple turnovers (**f**), PhnJ-Gly-d_2 in H_2O for a single-turnover experiment (**g**) and PhnJ-Gly-d_2 in H_2O for multiple turnovers (**h**). Typical reaction compositions were 150 μM PhnJ, 75 μM PRPn for single-turnover experiments, 1.5 mM PRPn for multiple-turnover experiments, 2 mM SAM, 1 mM dithionite, 50 units Factor Xa, 150 mM HEPES (pH 8.5), 0.5 M NaCl and 10% (w/v) glycerol. Typical reaction volume was 1.0 ml; typical headspace volume was 500 μl. a.u., arbitrary units.

conducted in 90% D_2O, the methane product was unlabelled with a mass-to-charge ratio (m/z) of 16 (Fig. 2e). When the reaction was initiated with ten enzyme equivalents of substrate in 90% D_2O, the methane product was predominantly labelled with deuterium with m/z 17 (Fig. 2f). When PhnJ-Gly-d_2 was used to initiate the reaction in H_2O with less than one enzyme equivalent of substrate, the isolated methane product predominantly contained a single deuterium label with m/z 17 (Fig. 2g). Finally, when PhnJ-Gly-d_2 was used to initiate the reaction in H_2O under multiple-turnover conditions, the methane product was unlabelled with m/z 16 (Fig. 2h). Under single-turnover conditions, the origin of the new hydrogen in the methane product derives exclusively from one of the Gly residues of PhnJ. Under multiple-turnover conditions, the origin of the new hydrogen in the methane product is determined from whether the reaction was conducted in

H_2O or D_2O. Therefore, during the course of the reaction catalysed by PhnJ, the active-site Gly residue directly participates in hydrogen atom transfer to the intermediate that forms methane. During multiple turnovers, the original hydrogen atoms contained within this Gly residue are ultimately replaced with those from bulk solvent.



(2)

According to the deuterium-labelling studies of PhnJ, one of the two prochiral hydrogen atoms from a PhnJ Gly residue is initially transferred to the transient Ado-CH_2• radical and the other hydrogen is transferred to the methyl group of the substrate during the course of the reaction. To determine the stereochemical origin of each of these hydrogen atom transfers, PhnJ was expressed in a medium containing Gly labelled with deuterium in the *pro-R* position and hydrogen in the *pro-S* position. The Gly used for this experiment was 68% *R*-[2-^{2}H]-Gly and 32% unlabelled Gly (Supplementary Fig. 4). Wild-type PhnJ was expressed in M9-minimal medium supplemented with 20 mM *R*-[2-^{2}H]-Gly. Mass spectrometric analysis of a tryptic peptide fragment, ^{27}AVAIPGYQVPFGGR^{40}, demonstrated that the average deuterium content at the *pro-R* position of the Gly residues in the isolated PhnJ (PhnJ-Gly-d_R) was ~52% (Supplementary Fig. 5). PhnJ-Gly-d_R was used to catalyse the C–P lyase reaction under conditions where the initial substrate concentration of PRPn was less than one equivalent of PhnJ. Under these single-turnover conditions, the Ado-CH_3 was shown by mass spectrometry to be 44% labelled with deuterium (Fig. 2d). No deuterium was found in the methane product (Supplementary Fig. 6). Therefore, the *pro-R* hydrogen of an unknown Gly from PhnJ is transferred to the Ado-CH_2• radical intermediate during the course of the C–P lyase reaction and the *pro-S* hydrogen is used in the formation of methane.

The identity of the specific Gly residue within PhnJ that is involved in two distinct hydrogen atom transfers during the course of the C–P lyase reaction was determined by two complementary experiments. In the first experiment, the reaction catalysed by PhnJ was conducted in 75% D_2O under multiple-turnover conditions. Under these reaction conditions, one of the Gly residues in PhnJ must exchange the *pro-R* and *pro-S* hydrogen atoms with deuterium from solvent. The reactions were quenched at various times and PhnJ was isolated. After proteolytic digestion with trypsin, the peptide fragments were analysed by mass spectrometry to identify those peptides that incorporated deuterium. The only peptide found labelled with deuterium during the course of this experiment was ^{27}VAIPGYQVPFGGR^{40} (Fig. 3a). After 12 min, the total deuterium content of a single Gly residue was ~36%. This peptide contains three Gly residues but only Gly 32 is absolutely conserved in PhnJ.

To confirm that Gly 32 is directly involved in hydrogen atom transfers to the transient Ado-CH_2• radical, we mutated this residue to alanine. The purified PhnJ Gly32Ala mutant was brown and it had the same absorbance maximum at 410 nm as wild-type PhnJ (Supplementary Fig. 3). After reduction with dithionite, the absorbance at 410 nm was lost, consistent with a redox-active [4Fe–4S]-cluster. When PhnJ Gly32Ala was incubated with all of the ingredients required for

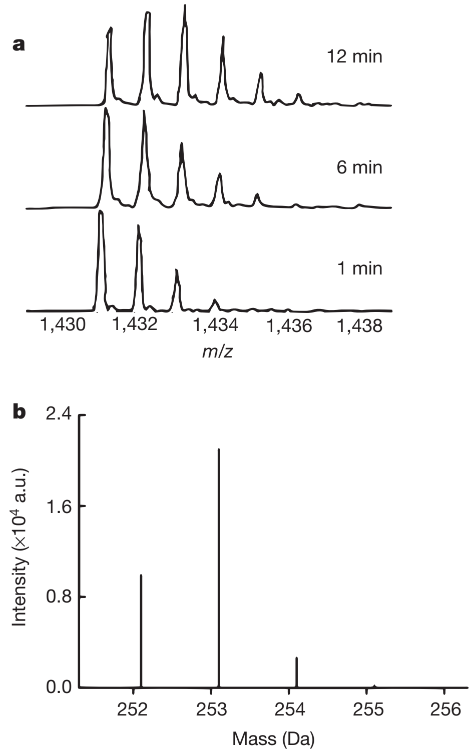


Figure 3 | Identification of Gly 32 as the site of the glycyl radical. **a**, Time course for the incorporation of deuterium into the peptide ^{27}VAIPGYQVPFGGR^{40} when the PhnJ reaction was conducted in D_2O. The reaction volume was 200 µl and contained 25 µM PhnJ, 300 µM PRPn, 1 mM SAM, 500 µM dithionite, 25 units Factor Xa, 150 mM HEPES (pH 8.5), 0.5 M NaCl and 10% (w/v) glycerol, in 75% D_2O. During the reaction course, 50 µl of reaction volume was loaded onto an SDS polyacrylamide gel and analysed by mass spectrometry after in-gel trypsin digestion. **b**, Mass spectrum of 5′-deoxyadenosine isolated from the PhnJ Gly32Ala mutant labelled with [2,3,3,3-$^2H_4$]-Ala. The reaction mixture contained 100 µM PhnJ-Gly32Ala-Ala-d_4, 1 mM PRPn, 2 mM SAM, 1 mM dithionite, 50 units Factor Xa, 150 mM HEPES (pH 8.5), 0.5 M NaCl and 10% (w/v) glycerol in a volume of 50 µl.

the C–P lyase reaction, the formation of PRcP and methane was not detected. However, Ado-CH_3 was detected and, thus, an Ado-CH_2• radical was formed in the active site of this mutant. To assess whether a hydrogen atom was abstracted by the Ado-CH_2• radical from Ala-32, we expressed the PhnJ Gly32Ala mutant in a minimal medium supplemented with L-[2,3,3,3-$^2H_4$]-Ala. The PhnJ Gly32Ala-Ala-d_4 mutant protein was incubated with the same ingredients described above for the unlabelled mutant and the Ado-CH_3 was isolated and subjected to mass spectrometric analysis; deuterated Ado-CH_3 (66%) was the major product (Fig. 3b). These results confirm that there is hydrogen atom transfer specifically from Gly 32 in PhnJ to the Ado-CH_2• radical during the reaction cycle.

On the basis of the experiments described in this report, we propose the following reaction mechanism for PhnJ during cleavage of the C–P bond in PRPn to form PRcP and methane (Fig. 4). PhnJ is a novel radical SAM enzyme that uses Gly 32 and Cys 272 during the cleavage of C–P bonds. In the proposed mechanism, the reaction starts with the reductive cleavage of SAM by the reduced [4Fe–4S]$^{1+}$-cluster to form the Ado-CH_2• radical intermediate. In the second step, the Ado-CH_2• intermediate abstracts the *pro-R* hydrogen from Gly 32 to generate Ado-CH_3 and a glycyl radical. In the third step, there is stereospecific hydrogen atom transfer from Cys 272 to the Gly 32 radical to make a thiyl radical on the side chain of Cys 272, and the Gly residue is regenerated. However, we note that there is no direct evidence for the formation of a thiyl radical in this study. In the fourth step, the thiyl radical attacks the phosphonate moiety of the substrate, PRPn, to create a transient thiophosphonate radical intermediate. Collapse of this intermediate, by

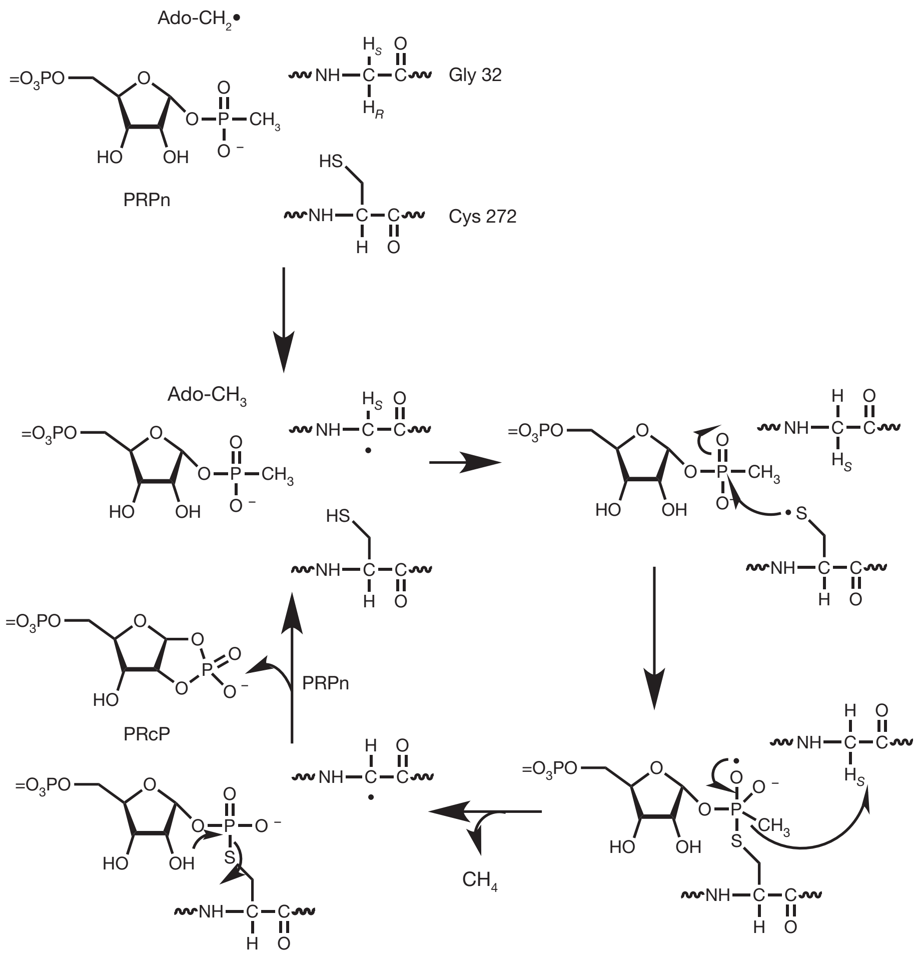
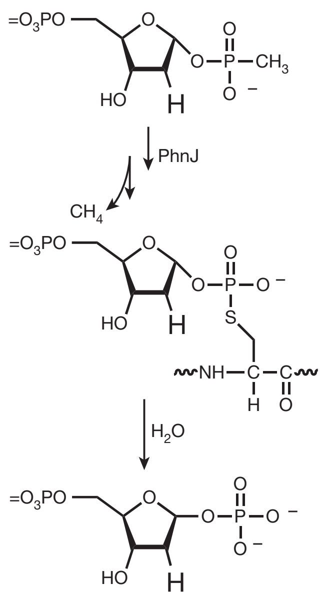


Figure 4 | Proposed mechanism for the reaction catalysed by PhnJ. The reaction cascade is initiated by formation of an Ado-CH2• radical. This intermediate abstracts the *pro-R* hydrogen from Gly 32 to form a glycyl radical. Hydrogen atom transfer from Cys 272 to the Gly 32 radical generates a thiyl radical on the side chain of Cys 272. This radical attacks the phosphonate moiety of the substrate to create a thiophosphonate radical intermediate. Homolytic C–P bond cleavage and hydrogen atom transfer from the original *pro-S* hydrogen of Gly 32 produces a thiophosphate intermediate, methane, and regenerates the radical intermediate at Gly 32. The ultimate product, PRcP, is formed by nucleophilic attack of the C2 hydroxyl on the covalent thiophosphate intermediate.

means of homolytic C–P bond cleavage and hydrogen atom transfer from the original *pro-S* hydrogen of Gly 32, produces a thiophosphate intermediate, methane, and regenerates the radical intermediate at Gly 32. The ultimate product, PRcP, is formed by nucleophilic attack of the C2 hydroxyl on the covalent thiophosphate intermediate. This reaction regenerates the free thiol group of Cys 272. After hydrogen atom transfer from Cys 272 to the Gly 32 radical, the entire process can be repeated with another substrate without the use of another molecule of SAM or any further involvement of the [4Fe–4S]-cluster.

The proposed reaction mechanism postulates the existence of a thiophosphate intermediate. To provide experimental support for this reaction intermediate, we synthesized the 2-deoxy substrate analogue, 2-dPRPn, in the anticipation that the lack of the C2 hydroxyl group of the substrate would trap the covalent intermediate as illustrated in equation (3)[8]. When PhnJ was incubated with ten enzyme equivalents of 2-dPRPn, and the other required ingredients of the C–P lyase reaction, ~2.3 enzyme equivalents of methane were detected by gas chromatography and approximately ~1.6 enzyme equivalents of 2-deoxyribose-1,5-bisphosphate were identified by ^{31}P-NMR spectroscopy (Supplementary Fig. 7). To assess whether any of the initial substrate was covalently attached to PhnJ, the enzyme was first treated with EDTA and then filtered through a 3 kDa membrane to remove the iron and all other small molecular weight molecules associated with the protein sample. The enzyme was digested with trypsin and the tryptic peptides analysed by ^{31}P-NMR to search for peptide fragments containing a phosphorylated enzyme-adduct. The ^{31}P-NMR spectrum revealed the appearance of two major resonances, one at a chemical shift of 4.5 p.p.m. and the other at a chemical shift of 25.0 p.p.m. (Fig. 5). The resonance at 4.5 p.p.m. correlates with the phosphate moiety at C5 of the proposed intermediate, and the resonance at 25.0 p.p.m. is consistent with the thiophosphate moiety at C1[20]. However, other types of phosphorus-containing adducts can resonate in this region of the ^{31}P-NMR spectrum and attempts to identify a modified tryptic peptide by mass spectrometry failed.



(3)

Pyruvate formate lyase and the anaerobic ribonucleotide reductase are two well-characterized proteins from the glycyl radical enzyme (GRE) family that utilize two subunits to catalyse their respective reactions[16–19]. Both of these systems have a smaller subunit, comprising ~250 amino acids, that functions as an activase (pyruvate formate lyase activase and ribonucleotide reductase activase). These proteins

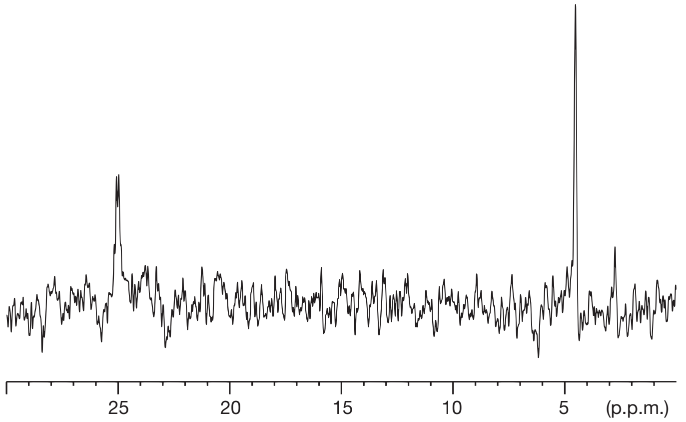


Figure 5 | ^{31}P-NMR spectrum of tryptic fragments of PhnJ after reaction with 2-dPRPn. In this experiment, 1 mM 2′-dPRPn was used as the substrate with 120 μM PhnJ. The reaction was quenched by the addition of EDTA and the PhnJ was isolated by ultrafiltration. PhnJ was digested with trypsin at pH 7.0. The resonance at 4.5 p.p.m. corresponds to the 5-phosphate of the ribose moiety of the covalent intermediate proposed in the mechanism depicted in equation (3). The resonance at 25 p.p.m. is consistent with the thiophosphate moiety of the proposed intermediate. The NMR spectrum was collected at pH 7.0 and 10 °C.

contain a radical SAM Cys motif, bind a redox-active [4Fe–4S]-cluster and generate a transient Ado-CH_2• intermediate from SAM. The Ado-CH_2• radical initiates the formation of a glycyl radical by abstracting the *pro-S* hydrogen of a highly conserved Gly residue on the larger subunit, of ~760 amino acids, which subsequently generates a catalytically competent thiyl radical at a conserved Cys residue of the larger subunit[21,22].

The polypeptide sequence of PhnJ consists of only 290 amino acids and is thus much smaller than other GREs (Supplementary Fig. 8). The hallmark of the proposed PhnJ reaction mechanism is the participation of a redox-active [4Fe–4S]-cluster, the transient formation of Ado-CH_2• from SAM, and the presence of two protein-based radicals from Gly 32 and Cys 272 that act in tandem for the cleavage of the C–P bond in phosphonate substrates. On the basis of labelling studies, the *pro-R* hydrogen of Gly 32 is abstracted by Ado-CH_2•, whereas the *pro-S* hydrogen is abstracted in all other GREs[21]. Both hydrogen atoms from Gly 32 are eventually transferred during the course of the C–P lyase reaction catalysed by PhnJ, which is unprecedented in any other GRE. The mechanistic characterization of the PhnJ reaction mechanism expands the repertoire of glycyl radical SAM enzymes and establishes a novel C–P bond cleaving reaction.

METHODS SUMMARY

Protein expression and purification. The gene for the expression of PhnJ was amplified and cloned into a pET42a(+) vector as an amino-terminal GST fusion protein, as described earlier[8]. For preparing PhnJ with deuterated Gly or Ala, the cells were grown in M9-minimal medium. The purification of wild-type PhnJ and mutant proteins was performed anaerobically in an MBraun LabMaster SP glove box, with oxygen levels of less than 4 p.p.m. The soluble protein fraction was applied to a GSTrap column (GE Healthcare, 5 ml) and eluted with reduced glutathione.

PhnJ-catalysed reactions. All PhnJ reactions were performed anaerobically (oxygen concentration less than 4 p.p.m. at all times) in an MBraun LabMaster SP glove box. A typical reaction contained 120–140 μM PhnJ, 2 mM SAM, 1 mM sodium dithionite, ×1 Factor Xa buffer, variable concentrations of PRPn, 150 mM HEPES buffer (pH 8.5), 0.5 M NaCl and 10% (w/v) glycerol. All reaction ingredients were incubated and the reaction initiated by the *in situ* cleavage of the GST tag by addition of 50 units of Factor Xa. Typical reaction volumes were 150–200 μl.

Received 15 November 2012; accepted 8 March 2013.
Published online 24 April 2013.

1. Reeburgh, W. S. Ocean methane biogeochemistry. *Chem. Rev.* **107,** 486–513 (2007).
2. Karl, D. M. *et al.* Aerobic production of methane in the sea. *Nature Geosci.* **1,** 473–478 (2008).
3. Metcalf, W. W. *et al.* Synthesis of methylphosphonic acid by marine microbes: a source of methane in the aerobic ocean. *Science* **337,** 1104–1107 (2012).
4. Wackett, L. P., Shames, S. L., Venditti, C. P. & Walsh, C. T. Bacterial carbon-phosphorus lyase: production, rates and regulation of phosphonic and phosphinic acid metabolism. *J. Bacteriol.* **169,** 710–717 (1987).
5. Frost, J. W., Loo, S., Cordiero, M. & Li, D. Radical-based dephosphorylation and organophosphonate biodegradation. *J. Am. Chem. Soc.* **109,** 2166–2171 (1987).
6. Wackett, L. P., Wanner, B. L., Venditti, C. P. & Walsh, C. T. Involvement of the phosphate regulon and the psiD locus in the carbon-phosphorus lyase activity of *Escherichia coli* K-12. *J. Bacteriol.* **169,** 1753–1756 (1987).
7. Metcalf, W. W. & Wanner, B. L. Mutational analysis of an *Escherichia coli* fourteen-gene operon for phosphonate degradation using Tn*phoA*′ elements. *J. Bacteriol.* **175,** 3430–3442 (1993).
8. Kamat, S. S., Williams, H. J. & Raushel, F. M. Intermediates in the transformation of phosphonates to phosphate by bacteria. *Nature* **480,** 570–573 (2011).
9. Cicchillo, R. M. *et al. Escherichia coli* lipoyl synthase binds two distinct [4Fe-4S] clusters per polypeptide. *Biochemistry* **43,** 11770–11781 (2004).
10. Cicchillo, R. M. *et al. Escherichia coli* quinolinate synthetase does indeed harbor a [4Fe-4S] cluster. *J. Am. Chem. Soc.* **127,** 7310–7311 (2005).
11. McGlynn, S. E. *et al.* Identification and characterization of a novel member of the radical AdoMet enzyme superfamily and implications for the biosynthesis of the Hmd hydrogenase active site cofactor. *J. Bacteriol.* **192,** 595–598 (2010).
12. Zhang, Y. *et al.* Diphthamide biosynthesis requires an organic radical generated by an iron-sulphur enzyme. *Nature* **465,** 891–896 (2010).
13. Sofia, H. J., Chen, G., Hetzler, B. G., Reyes-Spindola, J. F. & Miller, N. E. Radical SAM, a novel superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visual methods. *Nucleic Acids Res.* **29,** 1097–1106 (2001).
14. Frey, P. A., Hegeman, A. D. & Ruzicka, F. J. The radical SAM superfamily. *Crit. Rev. Biochem. Mol. Biol.* **43,** 63–88 (2008).
15. Booker, S. J. & Grove, T. L. Mechanistic and functional versatility of radical SAM enzymes. *F1000 Biol. Rep.* **2,** 52 (2010).
16. Eklund, H. & Fontecave, M. Glycyl radical enzymes: a conservative structural basis for radicals. *Structure* **7,** R257–R262 (1999).
17. Logan, D. T., Andersson, J., Sjoberg, B. M. & Nordlund, P. A glycyl radical site in the crystal structure of a class III ribonucleotide reductase. *Science* **283,** 1499–1504 (1999).
18. Becker, A. *et al.* Structure and mechanism of the glycyl radical enzyme pyruvate formate-lyase. *Nature Struct. Biol.* **6,** 969–975 (1999).
19. Vey, J. L. *et al.* Structural basis for glycyl radical formation by pyruvate formate-lyase activating enzyme. *Proc. Natl Acad. Sci. USA* **105,** 16137–16141 (2008).
20. Ghanem, E., Li, Y., Xu, C. & Raushel, F. M. Characterization of a phosphodiesterase capable of hydrolyzing EA 2192, the most toxic degradation product of the nerve agent VX. *Biochemistry* **46,** 9032–9040 (2007).
21. Frey, M., Rothe, M., Wagner, A. F. V. & Knappe, J. Adenosyl methionine-dependent synthesis of the glycyl radical in pyruvate formate-lyase by abstraction of the glycine C-2 pro-*S* hydrogen atom. *J. Biol. Chem.* **269,** 12432–12437 (1994).
22. Licht, S., Garfen, G. J. & Stubbe, J. Thiyl radicals in ribonucleotide reductases. *Science* **271,** 477–481 (1996).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Barondeau for use of the anaerobic chamber, R. Stipanovic for use of the gas chromatography mass spectrometer, A. Mehta and T. Begley for use of the liquid chromatography mass spectrometer, and C. Hilty for use of the ^{31}P-NMR spectrometer. We thank P. A. Lindahl for help with the EPR measurements (GM084266). This work was supported by the Robert A. Welch Foundation (A-840).

Author Contributions S.S.K., H.J.W. and F.M.R. designed the experiments. S.S.K. did the cloning and purification, performed the reactions and made all samples for analysis. S.S.K. and H.J.W. did the NMR, gas chromatography and gas chromatography mass spectrometry experiments. M.C. collected and analysed the EPR data. S.S.K. and L.J.D. did the trypsin digestion and peptide analysis. The manuscript was written by S.S.K. and F.M.R.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.M.R. (raushel@tamu.edu).

doi:10.1038/nature12120

Structure of active β-arrestin-1 bound to a G-protein-coupled receptor phosphopeptide

Arun K. Shukla[1]*, Aashish Manglik[2]*, Andrew C. Kruse[2]*, Kunhong Xiao[1], Rosana I. Reis[1], Wei-Chou Tseng[1], Dean P. Staus[1], Daniel Hilger[2], Serdar Uysal[3], Li-Yin Huang[1], Marcin Paduch[3], Prachi Tripathi-Shukla[1], Akiko Koide[3], Shohei Koide[3], William I. Weis[2,4], Anthony A. Kossiakoff[3], Brian K. Kobilka[2] & Robert J. Lefkowitz[1,5,6]

The functions of G-protein-coupled receptors (GPCRs) are primarily mediated and modulated by three families of proteins: the heterotrimeric G proteins, the G-protein-coupled receptor kinases (GRKs) and the arrestins[1]. G proteins mediate activation of second-messenger-generating enzymes and other effectors, GRKs phosphorylate activated receptors[2], and arrestins subsequently bind phosphorylated receptors and cause receptor desensitization[3]. Arrestins activated by interaction with phosphorylated receptors can also mediate G-protein-independent signalling by serving as adaptors to link receptors to numerous signalling pathways[4]. Despite their central role in regulation and signalling of GPCRs, a structural understanding of β-arrestin activation and interaction with GPCRs is still lacking. Here we report the crystal structure of β-arrestin-1 (also called arrestin-2) in complex with a fully phosphorylated 29-amino-acid carboxy-terminal peptide derived from the human V2 vasopressin receptor (V2Rpp). This peptide has previously been shown to functionally and conformationally activate β-arrestin-1 (ref. 5). To capture this active conformation, we used a conformationally selective synthetic antibody fragment (Fab30) that recognizes the phosphopeptide-activated state of β-arrestin-1. The structure of the β-arrestin-1–V2Rpp–Fab30 complex shows marked conformational differences in β-arrestin-1 compared to its inactive conformation. These include rotation of the amino- and carboxy-terminal domains relative to each other, and a major reorientation of the 'lariat loop' implicated in maintaining the inactive state of β-arrestin-1. These results reveal, at high resolution, a receptor-interacting interface on β-arrestin, and they indicate a potentially general molecular mechanism for activation of these multifunctional signalling and regulatory proteins.

Binding of β-arrestins to phosphorylated GPCRs is thought to involve two types of interaction between a receptor and a β-arrestin molecule[6]. A phosphate sensor engages the phosphorylated carboxy terminus or third intracellular loop of the receptor, and a conformational sensor recognizes the agonist-induced, active conformation of the core of the receptor (Fig. 1a). Using mass-spectrometry-based conformational mapping, we have previously used a V2 vasopressin-receptor-derived phosphopeptide (V2Rpp) to investigate activation of β-arrestin-1 and β-arrestin-2 (also known as arrestin-3)[5,7]. Binding to V2Rpp recapitulates functionalities of receptor-activated β-arrestins, such as enhanced clathrin binding[5]. Thus, we reasoned that crystallographic study of a complex of β-arrestin-1 with V2Rpp would provide insight into the mechanisms of receptor-mediated β-arrestin activation. However, well-ordered crystals of β-arrestin-1 bound to V2Rpp could not be obtained. This is presumably due to the significant conformational flexibility of activated arrestin molecules, as was recently determined for visual arrestin (also called arrestin-1) by NMR spectroscopy[8]. Given the success of antigen binding fragments (Fabs)[9] and nanobodies[10] in stabilizing particular GPCR conformations, we sought to identify and characterize conformationally selective Fabs that stabilize the V2Rpp bound, active conformation of β-arrestin-1.

We used a minimalist synthetic Fab phage display library[11] to select several high-affinity Fabs that selectively recognize the β-arrestin-1–V2Rpp complex (Supplementary Fig. 1). One of these, Fab30, displays marked selectivity for the activated conformation of β-arrestin-1 induced by V2Rpp (Fig. 1b). To ensure that Fab30 stabilizes a physiologically relevant conformation of β-arrestin-1, we investigated whether this Fab could facilitate interaction between a receptor and β-arrestin-1. Here, we used the previously described chimaeric receptor β_2-V2R which has an identical C terminus to V2Rpp, and which also has unaltered ligand-binding characteristics compared to the wild-type β_2 adrenergic receptor (β_2AR)[12]. Complexes of GPCRs with either G proteins or β-arrestins display an enhanced affinity for agonists due to the allosteric interactions among the agonist, the receptor and the transducer (G protein or β-arrestin)[13,14]. Addition of exogenous β-arrestin-1 to the membranes containing phosphorylated β_2-V2R resulted in a small fraction of the receptor in a high-agonist affinity state compared to receptor alone (Fig. 1c). Addition of Fab30 significantly increased the percentage of receptors in the high-affinity state. Furthermore, a direct physical stabilization of the receptor–β-arrestin-1 complex by Fab30 was revealed by co-immunoprecipitation (Fig. 1d). Here we present a 2.6 Å crystal structure of the β-arrestin-1–V2Rpp–Fab30 complex (Fig. 1e).

The overall structure of activated β-arrestin-1 exhibits a wide variety of pronounced structural changes compared to previously determined inactive state structures. Most notably, the N and C domains of β-arrestin-1 undergo a substantial twist relative to one another (Fig. 2a, b), with a 20° rotation around a central axis. The V2Rpp binds to the N domain at a similar location to the β-arrestin-1 C terminus in inactive structures and makes extensive contacts, primarily through charge–charge interactions of V2Rpp phosphates with β-arrestin-1 arginine and lysine side chains (compare Fig. 2c with Figs 2d and 3d).

This binding mode is consistent with previous limited proteolysis studies that revealed protection of the N domain of β-arrestin-1 in the presence of V2Rpp[5]. Additionally, crosslinking experiments on the β-arrestin-1–V2Rpp complex in the absence of Fab30 show that the N terminus of V2Rpp is in close proximity to K77, consistent with our structure (Supplementary Fig. 2). Like the β-arrestin-1 C terminus, V2Rpp binds β-arrestin-1 by extending the N-domain β-sandwich fold. Unlike the C terminus, however, V2Rpp binds as an antiparallel β-strand. This binding mode may serve as a general mechanism by which arrestins recognize the phosphorylated loops and C-terminal tails of receptors.

In addition to the large interdomain rearrangement, the N domain and central loops show large structural changes associated with

[1]Department of Medicine, Duke University Medical Center, Durham, North Carolina 27710, USA. [2]Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA. [3]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637, USA. [4]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305, USA. [5]Howard Hughes Medical Institute. Duke University Medical Center, Durham, North Carolina 27710, USA. [6]Department of Biochemistry, Duke University Medical Center, Durham,
*These authors contributed equally to this work.

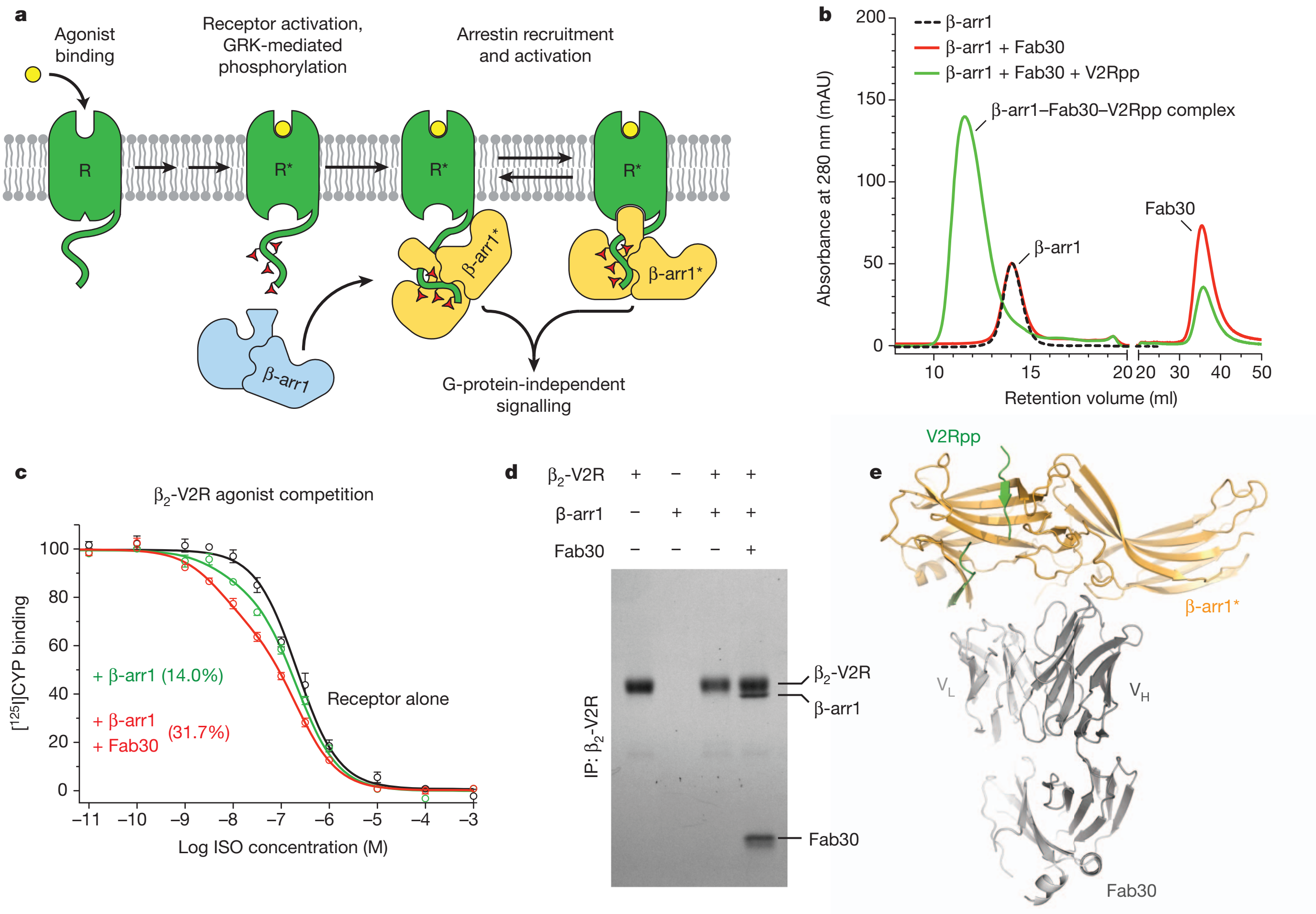


Figure 1 | Fab30 specifically recognizes and stabilizes an active state of β-arrestin-1. **a**, GPCRs are phosphorylated after activation, leading to the binding of arrestins. Interactions between the phosphorylated receptor and β-arrestin-1 lead to β-arrestin-1 activation and the subsequent blockade of G-protein signalling and initiation of β-arrestin-1 signalling pathways. **b**, Interaction between β-arrestin-1 and Fab30 requires the presence of V2Rpp in a size exclusion assay. **c**, The formation of a complex between a GPCR and β-arrestin allosterically leads to an enhanced affinity of agonist for the receptor, termed the 'high agonist affinity state'. Therefore, the fraction of receptor in the high-agonist affinity state reflects the extent of complex formation between receptor and β-arrestin. In a radioligand competition binding assay using ^{125}I-cyanopindolol as the probe and the agonist isoproterenol (Iso) as the competitor, β-arrestin-1 alone shifts a small portion (14%) of receptors into the high agonist affinity state. Fab30 significantly amplifies this effect (31%) ($n = 3$, $P < 0.0001$ in F test). **d**, In a pull-down assay, phosphorylated β_2–V2R chimaera shows appreciable binding to β-arrestin-1 only in the presence of Fab30. **e**, Overall structure of the β-arrestin-1–V2Rpp–Fab30 complex.

β-arrestin-1 activation. Several loops have been implicated in various aspects of β-arrestin activation and receptor interaction[15]. These include the 'finger loop' (residues 63–75), the 'middle loop' (residues 129–140) and the 'lariat loop' (residues 274–300). Each of these loops exhibits activation-dependent conformational changes (Fig. 2c–e). Comparison of these loops with inactive structures of β-arrestin-1 shows the considerable flexibility in each loop in the inactive conformation, but a more marked change in conformation upon β-arrestin activation (Fig. 2e). The crystal structure reveals that the V2Rpp occludes the inactive conformation of the finger loop, which has been shown to be important for arrestin discrimination between active and inactive GPCRs[16]. V2Rpp may stabilize an extended conformation of this loop to facilitate contact with the receptor core (Fig. 3a, b). It is noteworthy that the finger and middle loops above are not at the β-arrestin1–Fab30 interaction interface (Supplementary Fig. 3), and therefore, the conformational reorientation observed for these loops probably reflects activation-dependent changes in β-arrestin-1. However, finger loop residues 63–67 and lariat loop residues 285–287 engage in crystal lattice contacts (Supplementary Fig. 4), so some caution is warranted in the interpretation of conformational changes in these regions.

Two major sets of intramolecular interactions have been proposed to constrain arrestins in an inactive conformation: the three-element interaction and the polar-core interaction. The three-element interaction consists of interactions between β-strand I, α-helix I and the C terminus of arrestin[17]. Disruption of this interaction by mutagenesis yields arrestins that are partially phosphorylation-independent in their binding to receptor[17], suggesting a key role for this interaction network in recognizing phosphorylated receptors. The crystal structure of β-arrestin-1 shows that two well-conserved residues, K10 and K11 on β-strand I, make charge–charge interactions with phosphorylated residues pS363 and pS357 of V2Rpp (Fig. 3d). Indeed, mutagenesis of the corresponding lysines in visual arrestin significantly decreases binding to phosphorylated, active rhodopsin, suggesting that these residues serve as essential phosphate recognition elements[17]. Furthermore, previous limited proteolysis studies have indicated that the C terminus of both visual and β-arrestins is released upon activation as part of the disruption of the three-element interaction[5,18,19]. Consistent with this model, we observe that the β-arrestin-1 C terminus is displaced by the V2Rpp (Fig. 3c, d). The β-arrestin-1 C terminus contains a clathrin binding site that has been previously characterized to be important for GPCR internalization[20]. Hence, displacement of the C terminus upon phosphopeptide binding and β-arrestin-1 activation is probably an important contributor to clathrin-mediated GPCR internalization. In comparison to previous models, however, binding of

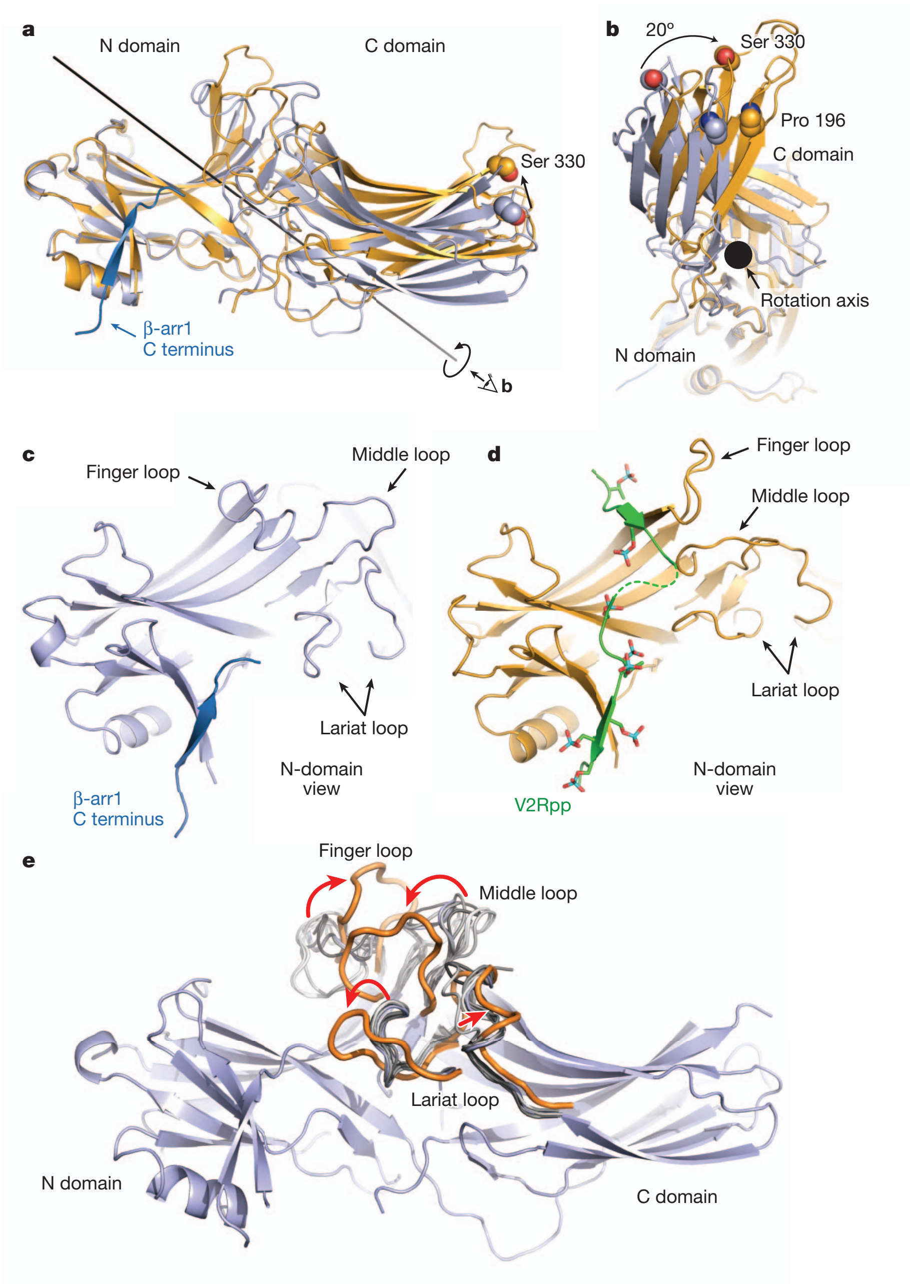


Figure 2 | Conformational changes associated with β-arrestin-1 activation. The structures of inactive β-arrestin-1 (Protein Data Bank (PDB) accession 1G4M; chain A, light blue) and active β-arrestin-1 (gold) were aligned on the N domains. The β-arrestin-1 C terminus is highlighted in dark blue. **a**, A substantial rotation and translation of the C domain relative to the N domain occurs upon activation. The rotation axis is indicated as a solid black line. **b**, View of C-domain rotation along the axis. **c**, N domain of inactive arrestin, highlighting important regions. **d**, Active β-arrestin-1 in the same orientation, showing V2Rpp in green. Phosphorylated residues are highlighted as sticks. **e**, The overall structure of inactive β-arrestin-1 (PDB 1G4M; chain A), with loops from all inactive β-arrestin-1 structures superimposed (grey loops). The active conformation of these loops (orange loops) deviates from all inactive structures.

V2Rpp and displacement of the C terminus does not markedly alter the secondary structure of β-strand I. Although we observe abundant charge–charge interactions between β-arrestin-1 and V2Rpp, it is noteworthy that neither the specific sequence of the phosphorylation sites nor the net number of phosphates is conserved among various receptors. Therefore, it remains to be seen how β-arrestins fine-tune their interaction with such a large number of receptors.

The second constraint that stabilizes the inactive conformation of arrestins is the polar core[18], consisting of five interacting charged residues: D26, R169, D290, D297 and R393. Disruption of the polar core by mutagenesis yields phosphorylation-independent mutants of both visual arrestin and β-arrestin-1 (refs 17, 21). Charge reversal of R169 or D290 in β-arrestin-1 (R175 and D296 in visual arrestin) disrupts this interaction network, yielding arrestins that can bind non-phosphorylated, activated receptors. On the basis of these studies, R169 was previously proposed to be a critical phosphate sensor in β-arrestin-1, and disruption of the polar core was proposed to be required for β-arrestin-1 activation[21]. Contrary to this model, R169 does not make any direct contacts with V2Rpp phosphates, indicating that direct interaction between R169 and receptor phosphates is not required for arrestin activation (Fig. 3e, f). However, binding of V2Rpp does disrupt the polar core. V2Rpp binding to β-arrestin-1 displaces the arrestin C terminus, and in doing so, removes R393 from the polar core. Residues D290 and D297 also lose interactions within the polar core, and this is accompanied by a marked twisting of the lariat loop, which contains both D290 and D297. Therefore, it is possible that the disruption of the polar core is driven by the excess negative charge in this region following displacement of the arrestin C terminus residue R393. Notably, the side chain of K294, a residue within the lariat loop, flips towards the N domain upon activation and engages pT360. It is possible that K294 recognition of phosphates provides an additional driving force for lariat-loop rearrangement, and may therefore stabilize β-arrestin-1 in an active conformation. This observation in the crystal structure is consistent with crosslinking experiments, which

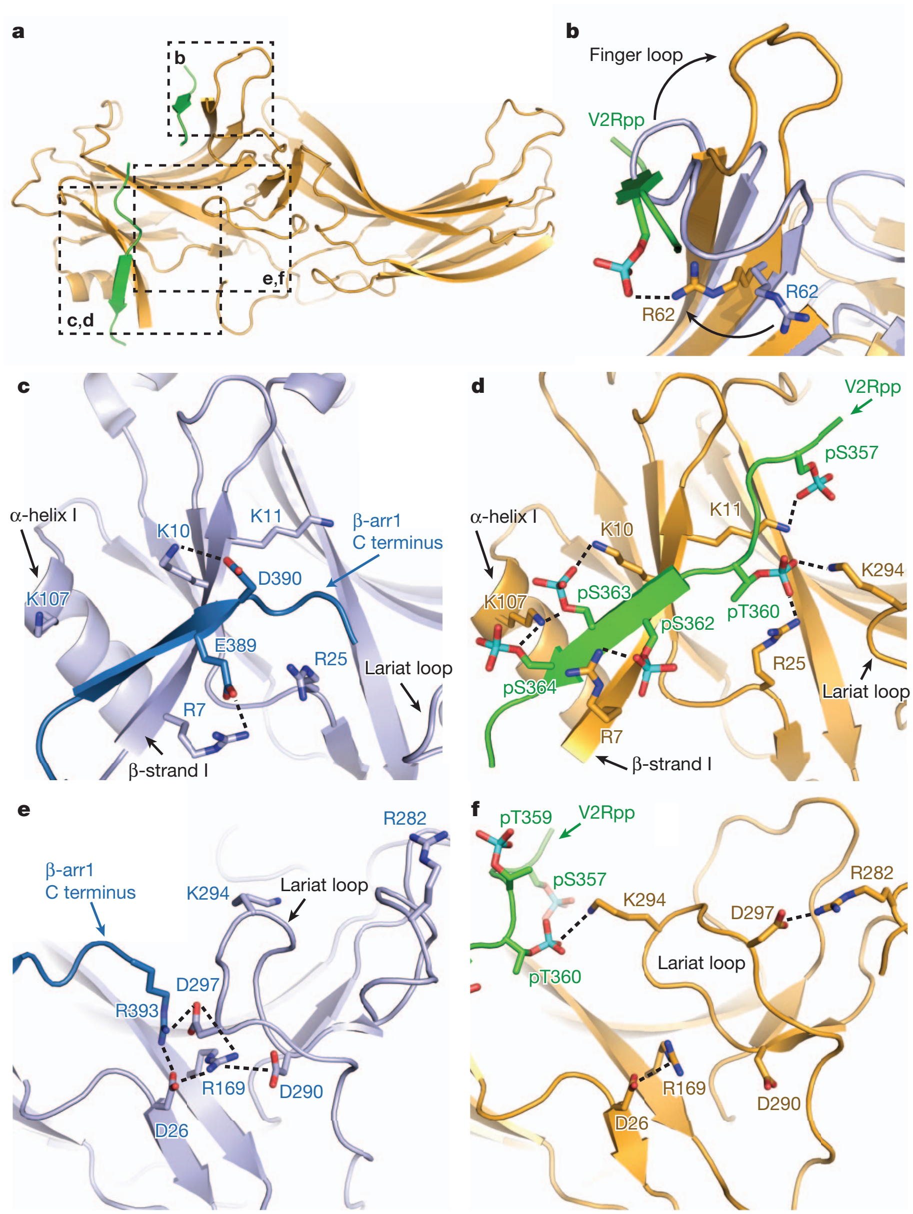


Figure 3 | V2Rpp interactions with β-arrestin-1. a, Overall view of β-arrestin-1, with regions of interest in boxes. **b**, V2Rpp (green) displaces the inactive finger loop (light blue), causing it to adopt an extended conformation in the active state (gold). Select charge–charge contacts are shown with dotted lines in **b**–**f**. **c**, In the inactive conformation, the β-arrestin-1 C-terminal β-strand (dark blue) lies along the N domain in the three-element interaction network. **d**, Upon activation, this strand is displaced by the C terminus of V2Rpp, which engages in extensive charge–charge interactions through phosphorylated residues. **e**, The polar core of β-arrestin-1 is thought to be a critical stabilizer of the inactive state. **f**, Upon V2Rpp binding, the C-terminal strand residue Arg 393 is displaced, and its interaction partner D297 undergoes a large movement together with the rest of the lariat loop.

reveal the disappearance of an intrapeptide crosslink between K292 and K294 in the presence of V2Rpp (Supplementary Fig. 5), indicating that V2Rpp induces a conformation like that seen in the crystal structure even in the absence of Fab30.

Whereas domain rearrangement upon arrestin activation has been proposed previously, the observed 20° twisting of the N and C domains of β-arrestin-1 upon activation is unanticipated. Biochemical studies have shown that sequential deletion of the visual arrestin hinge region connecting the N and C domains results in a progressive decrease in the ability of arrestin to bind phosphorylated, light-activated rhodopsin. This suggests a requirement for relative movement of the two domains for efficient interaction with activated receptors[22]. However, the twisting motion observed here stands in contrast to the 'clamshell' hypothesis advanced previously[23]. Considering the large number of interaction partners of β-arrestins during cellular signalling[24], it is tempting to speculate that the twisting movement of the two domains upon arrestin activation may expose interaction interfaces with such binding partners.

Recent NMR and double electron–electron resonance (DEER) studies have assessed the conformational changes induced in visual arrestin upon interaction with phosphorylated, light-activated rhodopsin[8,25]. Intriguingly, NMR spectroscopy of activated visual arrestin revealed significant line broadening attributed to intermediate timescale conformational dynamics over the entire arrestin molecule[8]. Within such an ensemble of activated arrestin conformations, Fab30 probably stabilizes a conformation of β-arrestin-1 that preferentially binds activated GPCRs. Furthermore, distance restraints for activated visual arrestin derived from DEER experiments are highly consistent with the active structure of β-arrestin-1 presented here (Supplementary Fig. 6). Most notably, the large conformational change observed for the middle loop by DEER spectroscopy upon binding light-activated,

phosphorylated rhodopsin is also evident in the crystal structure of activated β-arrestin-1. Given the importance of this region in maintaining the inactive conformation of visual arrestin, the agreement in conformational changes within arrestin suggests that the V2Rpp-bound, active conformation of β-arrestin-1 presented here represents a similar state to that of arrestin in complex with a phosphorylated, activated GPCR. This further suggests that the conformational changes associated with activation and receptor binding are conserved throughout the arrestin family. However, the binding stoichiometry between GPCRs and arrestins still remains to be fully established. Recent biochemical studies have suggested that two rhodopsin molecules may simultaneously bind one arrestin[26]. The extensive and specific contacts between V2Rpp and the β-arrestin-1 N domain probably preclude another receptor C terminus from binding β-arrestin-1. However, it is possible that an arrestin molecule bound to the phosphorylated C terminus of a receptor could interact with the seven-transmembrane-segment core of another receptor. Additional data, including a crystal structure of a GPCR–β-arrestin complex, will be required to clarify this.

In summary, we present here the structure of an activated arrestin bound to the phosphorylated C terminus of a GPCR. The structure not only provides the atomic details of a potentially general GPCR–β-arrestin interaction interface, but also offers novel insights into the activation process of arrestins, and reveals a large interdomain twisting associated with activation. These findings will facilitate future efforts to understand the structural basis for β-arrestin activation and signalling. Such studies may ultimately yield insight into how GPCRs achieve such a large breadth of signalling complexity.

METHODS SUMMARY

A truncated version of β-arrestin-1 containing residues 1–393 was expressed in *Escherichia coli* strain BL21(DE3) and purified to homogeneity using a GST tag and anion exchange chromatography. Fab30 was expressed in *E. coli* strain 55244 and purified by protein A and cation exchange chromatography. Mixture and incubation of the components yielded the β-arrestin-1–V2Rpp–Fab30 complex, which was purified by size-exclusion chromatography and crystallized using vapour diffusion. Diffraction data were collected at GM/CA-CAT beamline 23ID-D at the Advanced Photon Source at Argonne National Laboratory.

Full Methods and any associated references are available in the online version of the paper.

Received 7 January; accepted 28 March 2013.
Published online 21 April 2013.

1. Pierce, K. L., Premont, R. T. & Lefkowitz, R. J. Seven-transmembrane receptors. *Nature Rev. Mol. Cell Biol.* **3,** 639–650 (2002).
2. Hepler, J. R. & Gilman, A. G. G proteins. *Trends Biochem. Sci.* **17,** 383–387 (1992).
3. Freedman, N. J. & Lefkowitz, R. J. Desensitization of G protein-coupled receptors. *Recent Prog. Horm. Res.* 51, 319–351; discussion 352–313 (1996).
4. Lefkowitz, R. J. & Shenoy, S. K. Transduction of receptor signals by β-arrestins. *Science* **308,** 512–517 (2005).
5. Nobles, K. N., Guan, Z., Xiao, K., Oas, T. G. & Lefkowitz, R. J. The active conformation of β-arrestin1: direct evidence for the phosphate sensor in the N-domain and conformational differences in the active states of β-arrestins1 and -2. *J. Biol. Chem.* **282,** 21370–21381 (2007).
6. Gurevich, V. V. & Gurevich, E. V. The structural basis of arrestin-mediated regulation of G-protein-coupled receptors. *Pharmacol. Ther.* **110,** 465–502 (2006).
7. Xiao, K., Shenoy, S. K., Nobles, K. & Lefkowitz, R. J. Activation-dependent conformational changes in β-arrestin 2. *J. Biol. Chem.* **279,** 55744–55753 (2004).
8. Zhuang, T. *et al.* Involvement of distinct arrestin-1 elements in binding to different functional forms of rhodopsin. *Proc. Natl Acad. Sci. USA* **110,** 942–947 (2013).
9. Rasmussen, S. G. F. *et al.* Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450,** 383–387 (2007).
10. Rasmussen, S. G. F. *et al.* Structure of a nanobody-stabilized active state of the β_2 adrenoceptor. *Nature* **469,** 175–180 (2011).
11. Fellouse, F. A. *et al.* High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J. Mol. Biol.* **373,** 924–940 (2007).
12. Oakley, R. H., Laporte, S. A., Holt, J. A., Barak, L. S. & Caron, M. G. Association of β-arrestin with G protein-coupled receptors during clathrin-mediated endocytosis dictates the profile of receptor resensitization. *J. Biol. Chem.* **274,** 32248–32257 (1999).
13. De Lean, A., Stadel, J. M. & Lefkowitz, R. J. A ternary complex model explains the agonist-specific binding properties of the adenylate cyclase-coupled β-adrenergic receptor. *J. Biol. Chem.* **255,** 7108–7117 (1980).
14. Gurevich, V. V., Pals-Rylaarsdam, R., Benovic, J. L., Hosey, M. M. & Onorato, J. J. Agonist-receptor-arrestin, an alternative ternary complex with high agonist affinity. *J. Biol. Chem.* **272,** 28849–28852 (1997).
15. Han, M., Gurevich, V. V., Vishnivetskiy, S. A., Sigler, P. B. & Schubert, C. Crystal structure of β-arrestin at 1.9 Å: possible mechanism of receptor binding and membrane translocation. *Structure* **9,** 869–880 (2001).
16. Hanson, S. M. *et al.* Differential interaction of spin-labeled arrestin with inactive and active phosphorhodopsin. *Proc. Natl Acad. Sci. USA* **103,** 4900–4905 (2006).
17. Vishnivetskiy, S. A. *et al.* An additional phosphate-binding element in arrestin molecule. Implications for the mechanism of arrestin activation. *J. Biol. Chem.* **275,** 41049–41057 (2000).
18. Vishnivetskiy, S. A. *et al.* How does arrestin respond to the phosphorylated state of rhodopsin? *J. Biol. Chem.* **274,** 11451–11454 (1999).
19. Palczewski, K., Buczylko, J., Imami, N. R., McDowell, J. H. & Hargrave, P. A. Role of the carboxyl-terminal region of arrestin in binding to phosphorylated rhodopsin. *J. Biol. Chem.* **266,** 15334–15339 (1991).
20. Goodman, O. B. Jr *et al.* β-arrestin acts as a clathrin adaptor in endocytosis of the β2-adrenergic receptor. *Nature* **383,** 447–450 (1996).
21. Kovoor, A., Celver, J., Abdryashitov, R. I., Chavkin, C. & Gurevich, V. V. Targeted construction of phosphorylation-independent beta-arrestin mutants with constitutive activity in cells. *J. Biol. Chem.* **274,** 6831–6834 (1999).
22. Vishnivetskiy, S. A., Hirsch, J. A., Velez, M. G., Gurevich, Y. V. & Gurevich, V. V. Transition of arrestin into the active receptor-binding state requires an extended interdomain hinge. *J. Biol. Chem.* **277,** 43961–43967 (2002).
23. Gurevich, V. V. & Gurevich, E. V. The molecular acrobatics of arrestin activation. *Trends Pharmacol. Sci.* **25,** 105–111 (2004).
24. Xiao, K. *et al.* Functional specialization of β-arrestin interactions revealed by proteomic analysis. *Proc. Natl Acad. Sci. USA* **104,** 12011–12016 (2007).
25. Kim, M. *et al.* Conformation of receptor-bound visual arrestin. *Proc. Natl Acad. Sci. USA* **109,** 18407–18412 (2012).
26. Sommer, M. E., Hofmann, K. P. & Heck, M. Distinct loops in arrestin differentially regulate ligand binding within the GPCR opsin. *Nature Commun.* **3,** 995 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Capel for technical assistance and V. Ronk, D. Addison and Q. Lennon for administrative and secretarial support. We thank S. Ahn and L. Wingler for critical reading of the manuscript. We acknowledge support from the Stanford Medical Scientist Training Program and the American Heart Association (A.M.), from the National Science Foundation (A.C.K.), from the National Institutes of Health Grants NS028471 (B.K.K.), HL16037 and HL70631 (R.J.L.), GM072688 and GM087519 (A.A.K. and S.K.), HL 075443 (K.X.) and from the Mathers Foundation (B.K.K. and W.I.W.). R.I.R is supported by a post-doctoral fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–CAPES. R.J.L. is an investigator with the Howard Hughes Medical Institute.

Author Contributions A.K.S. conceived the project, designed the Fab selection strategy, selected and characterized Fab30, established and optimized complex formation and purification conditions, prepared protein for crystallization trials and supervised the experiments related to the biochemical characterization of the complex. A.M. purified the complex, performed crystallography trials and grew crystals. A.M. and A.C.K. collected and processed diffraction data, and solved and refined the structure with supervision from W.I.W. R.I.R. assisted with advanced Fab characterization and optimized complex formation. W.-C.T. assisted with Fab selection and preliminary characterization. K.X. performed and analysed the crosslinking experiments. D.P.S. performed and analysed radioligand binding experiments. L.-Y.H. assisted with functional characterization of the complex. P.T.-S. expressed and purified the receptor. S.U., M.P., A.K., S.K. and A.A.K. generated and provided the phage display library and the screening protocol and helped with the initial phase of Fab selection. D.H. performed the comparison of the structural model with EPR data. A.K.S., A.M. and A.C.K. made figures. A.K.S., A.M., A.C.K., B.K.K. and R.J.L. wrote the manuscript. B.K.K. and R.J.L. supervised the overall research.

Author Information Coordinates and structure factors for the β-arrestin-1–V2Rpp–Fab30 complex are deposited in the Protein Data Bank under accession code 4JQI. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.K.K. (kobilka@stanford.edu) or R.J.L. (lefko001@receptor-biol.duke.edu).

METHODS

Purification of β-arrestin-1. Full-length β-arrestin-1 was purified from *E. coli* as described previously[5]. Briefly, GST-tagged rat β-arrestin1 in the pGEX4T vector was transformed into BL21(DE3) cells, large-scale expression cultures were grown in Terrific broth, and induced with 1 mM IPTG for 16 h at 16 °C. Cell pellets were lysed in 20 mM HEPES pH 7.4, 150 mM NaCl, 1 mM PMSF and 2 mM dithiothreitol (DTT) using a micro fluidizer, and the lysate was bound to glutathione sepharose at 4 °C for 2 h. Beads were washed in lysis buffer and β-arrestin-1 was eluted by overnight incubation with thrombin at 4 °C. β-arrestin-1 was then purified with a HiTrap Q column and eluted by a linear gradient of NaCl. Peak fractions were pooled and purified protein was dialysed in 20 mM HEPES pH 7.4 and 150 mM NaCl.

Selection and characterization of Fab. The phage library was panned against biotinylated β-arrestin-1 bound to V2Rpp and immobilized on streptavidin beads. Beads were washed three times and bound phages were amplified by infecting *E. coli* XL-1 blue cells. Amplified phage were precipitated and used for a second and third round of panning. To select against Fabs that bind to the inactive conformation of β-arrestin-1, beads coated with the β-arrestin-1–V2Rpp complex were first incubated with phage and then with 1 μM non-biotinylated β-arrestin-1. Subsequently, phage were eluted with dithiothreitol (DTT) and resulting clones were used for single point ELISA to test their selectivity towards β-arrestin-1 bound to V2Rpp. ELISA positive clones were sequenced and further characterized.

Selectivity of Fabs towards V2Rpp-bound β-arrestin-1 conformation. β-arrestin-1 was incubated with either non-phosphorylated V2 vasopressin peptide (V2Rnp) or V2Rpp in a 1:3 molar ratio for 30 min at 25 °C. Subsequently, purified Fabs were added at a 1:2 molar ratio with β-arrestin-1 and incubated for additional 30 min at 25 °C. Then, pre-washed Protein A beads (Pierce) were added to the reactions and incubated for 30 min at 25 °C. The final concentration of β-arrestin-1 in the binding reaction was 10 nM. Beads were washed four times with 1 ml buffer (20 mM HEPES pH 7.4, 150 mM NaCl) and proteins were eluted with SDS–PAGE gel loading buffer. The eluted proteins were run on a 4–20% SDS–PAGE gel. Fab30 displayed the greatest difference in its ability to co-immunoprecipitate β-arrestin-1 between V2Rnp and V2Rpp and was therefore chosen for further characterization (R.I.R. and R.J.L., manuscript in preparation).

Radioligand binding. Sf9 insect cells were co-infected with baculovirus encoding an N-terminal Flag-tagged β_2-V2R (a chimaeric receptor with β_2AR residues 1–341 and V2 vasopressin receptor residues 328–372) and GRK2-CAAX (GRK2 with a membrane tethering prenylation signal). After viral infection for 72 h at 27 °C, cells were incubated with 10 μM isoproterenol at 37 °C for 15 min to induce receptor phosphorylation. Subsequently, the cells were washed and membranes were prepared and flash frozen. Membranes were extensively washed to remove isoproterenol used for receptor phosphorylation. For radioligand binding, membranes were incubated with 60 pM [^{125}I]cyanopindolol (GE Healthcare Life-science) in radioligand binding buffer (50 mM Tris, pH 7.4, 50 mM potassium acetate, 0.5 mM magnesium chloride, 1 mM ascorbic acid) with varying concentrations of freshly prepared isoproterenol. Binding reactions were performed in parallel, with 1 μM β-arrestin-1 (residues 1–393) incubated either in the presence or absence of 10 μM Fab30. Binding reactions were incubated for 90 min at 27 °C, followed by rapid harvesting on a GF-B filter and scintillation counting in a Packard gamma counter. Competition binding data were analysed by a nonlinear curve-fitting procedure where low- and high-affinity values were computed globally using a two-site binding model (GraphPad Prism). The *F*-test was used to test whether Fab30 significantly altered the amount of β_2-V2R coupled to β-arrestin.

Effect of Fab30 on β_2-V2R–β-arrestin-1 interaction. Fab30 was expressed and purified as described previously[27]. β_2-V2R was expressed in Sf9 cells and purified as described previously[28]. Purified, phosphorylated β_2-V2R was prepared bound to the potent β_2AR agonist BI-167107[10] and incubated at a concentration of 1 μM with 3 μM β-arrestin-1 with and without Fab30 at 25 °C for 2 h in a buffer comprised of 20 mM HEPES, pH 7.4, 150 mM NaCl, 0.01% MNG (lauryl maltose neopentyl glycol). Subsequently, β_2-V2R was immunoprecipitated using M1 Flag antibody beads. Beads were washed and protein was eluted with 5 mM EDTA and 0.25 mg ml^{-1} Flag peptide and elution fractions were analysed on a 4–20% SDS–PAGE gel and stained with Coomassie.

Preparation of β-arrestin-1–V2Rpp–Fab30, crystallization and structure determination. β-arrestin-1 (20 μM) was incubated with V2Rpp (27 μM) for 30 min at 25 °C. An excess of Fab30 was added and the complex was incubated for 1 h at 25 °C. The β-arrestin-1–V2Rpp–Fab30 complex was purified from excess Fab30 and V2Rpp by size exclusion chromatography in 20 mM HEPES pH 7.5, 150 mM NaCl and 1 mM TCEP. The purified complex was concentrated to 8 mg ml^{-1} using a centrifugal concentrator (Vivaspin, GE Healthcare). Crystals were grown in hanging drops containing 1 μl of complex solution and 0.5 μl of a well solution composed of 17% PEG 3350, 0.1 M HEPES pH 7.5, and 0.2 M L-proline. Drops were stored at 20 °C and crystals appeared within 24 h and grew to full size within 3 days (Supplementary Fig. 4). Crystals were flash frozen in liquid nitrogen after a 30 s soak in 19% PEG 3350, 0.1 M HEPES pH 7.5, 0.2 M L-proline, and 20% ethylene glycol.

Diffraction data were collected at the Advanced Photon Source GM/CA-CAT beamline 23ID-D. Although typical crystals grew to over 300 μm in two dimensions, and over 100 μm in the third dimension, we used a 10-μm-sized beam to collect multiple full data sets from the highest quality regions of the crystal. A full data set from the single best region of the crystal was indexed, integrated and scaled with HKL-2000[29]. The structure of the complex was solved by molecular replacement using Phaser[30]. Owing to the conformational changes observed for β-arrestin-1, it proved necessary to first search for Fab30 (PDB 3EFF; Fab2, with the complementary determining regions omitted, was used as a search model for Fab30)[31], followed by only the C domain of β-arrestin-1 (PDB 1JSY)[32]. A subsequent search for the N domain of β-arrestin-1 failed in multiple attempts; the N domain was then manually placed to fit the electron density. A significant decrease in R_{free} upon rigid-body refinement of the N domain provided confidence in the final molecular replacement solution. The resulting model was then iteratively refined by building regions of β-arrestin-1, V2Rpp and Fab30 in Coot[33] and refining in Phenix[34]. We used translation libration screw-motion (TLS) refinement with groups defined within Phenix[34]. For the V2Rpp, the amino-acid register was determined by the strong electron density resulting from electron-rich phosphates on phosphoserine and phosphothreonine residues. As shown in Supplementary Fig. 7, the electron density for V2Rpp was clear, permitting confident placement of most side chains. We used MolProbity[35] to assess statistics for the final model of the β-arrestin-1–V2Rpp–Fab30 complex. Supplementary Table 1 outlines statistics for data collection and refinement. Figures were prepared in PyMOL[36] and secondary structure was assigned using the DSSP algorithm[37]. Domain rotation was measured and analysed with DynDom[38].

27. Rizk, S. S. *et al.* Allosteric control of ligand-binding affinity using engineered conformation-specific effector proteins. *Nature Struct. Mol. Biol.* **18,** 437–442 (2011).
28. Kobilka, B. K. Amino and carboxyl terminal modifications to facilitate the production and purification of a G protein-coupled receptor. *Anal. Biochem.* **231,** 269–271 (1995).
29. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Macromol. Crystallogr. A* **276,** 307–326 (1997).
30. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
31. Uysal, S. *et al.* Crystal structure of full-length KcsA in its closed conformation. *Proc. Natl Acad. Sci. USA* **106,** 6644–6649 (2009).
32. Milano, S. K., Pace, H. C., Kim, Y. M., Brenner, C. & Benovic, J. L. Scaffolding functions of arrestin-2 revealed by crystal structure and mutagenesis. *Biochemistry* **41,** 3321–3328 (2002).
33. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
34. Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Crystallogr. D* **61,** 850–855 (2005).
35. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
36. Schrodinger, L. The PyMOL Molecular Graphics System v.1.3r1 (2010).
37. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).
38. Hayward, S., Kitao, A. & Berendsen, H. J. Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins* **27,** 425–437 (1997) CrossRef.

doi:10.1038/nature12133

Crystal structure of pre-activated arrestin p44

Yong Ju Kim[1], Klaus Peter Hofmann[1,2], Oliver P. Ernst[3], Patrick Scheerer[4], Hui-Woog Choe[1,5] & Martha E. Sommer[1]

Arrestins interact with G-protein-coupled receptors (GPCRs) to block interaction with G proteins[1,2] and initiate G-protein-independent signalling[3]. Arrestins have a bi-lobed structure that is stabilized by a long carboxy-terminal tail (C-tail), and displacement of the C-tail by receptor-attached phosphates activates arrestins for binding active GPCRs[4]. Structures of the inactive state of arrestin are available[5,6], but it is not known how C-tail displacement activates arrestin for receptor coupling. Here we present a 3.0 Å crystal structure of the bovine arrestin-1 splice variant p44, in which the activation step is mimicked by C-tail truncation. The structure of this pre-activated arrestin is profoundly different from the basal state and gives insight into the activation mechanism. p44 displays breakage of the central polar core and other interlobe hydrogen-bond networks, leading to a 21° rotation of the two lobes as compared to basal arrestin-1. Rearrangements in key receptor-binding loops in the central crest region include the finger loop[7–9], loop 139 (refs 8, 10, 11) and the sequence Asp 296–Asn 305 (or gate loop), here identified as controlling the polar core. We verified the role of these conformational alterations in arrestin activation and receptor binding by site-directed fluorescence spectroscopy. The data indicate a mechanism for arrestin activation in which C-tail displacement releases critical central-crest loops from restricted to extended receptor-interacting conformations. In parallel, increased flexibility between the two lobes facilitates a proper fitting of arrestin to the active receptor surface. Our results provide a snapshot of an arrestin ready to bind the active receptor, and give an insight into the role of naturally occurring truncated arrestins in the visual system.

Arrestin binding to the active, phosphorylated receptor (R*P) is a multistep process[12]. Basal arrestin first interacts with receptor-attached phosphates in an initial complex (that is, pre-binding)[13,14], which displaces the C-tail. This activation step primes arrestin for an intramolecular conversion[15] that brings about the high-affinity arrestin–R*P complex. For this study, we prepared C-terminally truncated bovine arrestin-1 (rod arrestin) for X-ray structure analysis. This truncated arrestin, in which the last 35 amino acids are replaced by a single alanine residue[16], occurs naturally as a splice variant called p44 and exists in a pre-activated state[13]. Recombinant p44 was expressed in *Escherichia coli* and purified similarly as full-length arrestin-1. We verified the constitutive activity of our p44 preparation by the 'extra Meta II' assay, which measures the binding of arrestin or p44 to the light-activated GPCR rhodopsin as a stabilization of the 380-nm-absorbing metarhodopsin II (Meta II). Like native p44, recombinant p44 stabilized Meta II in both phosphorylated and non-phosphorylated receptor preparations and had a higher affinity for light-activated phosphorylated rhodopsin (R*P) than arrestin-1[17] (Supplementary Fig. 1). Purified p44, crystallized in the presence of the apo-receptor opsin[18,19] (see Methods), yielded crystals in the hexagonal $P6_122$ space group, which no previous arrestin or p44 crystal has adopted. (For data collection, structure determination and refinement statistics, see Methods and Supplementary Table 1.) The current p44 structure displays several significant conformational differences to the parent protein as seen in superposition of basal arrestin-1 and p44 (Fig. 1) and from a comparison of the secondary structures of arrestin-1 and p44 (Supplementary Fig. 2).

One of the most functionally important differences between the current p44 structure and that of basal arrestin-1 is the state of the polar core, a buried hydrogen-bond network comprised of residues from both N and C domains and the C-tail, where Arg 175 forms a

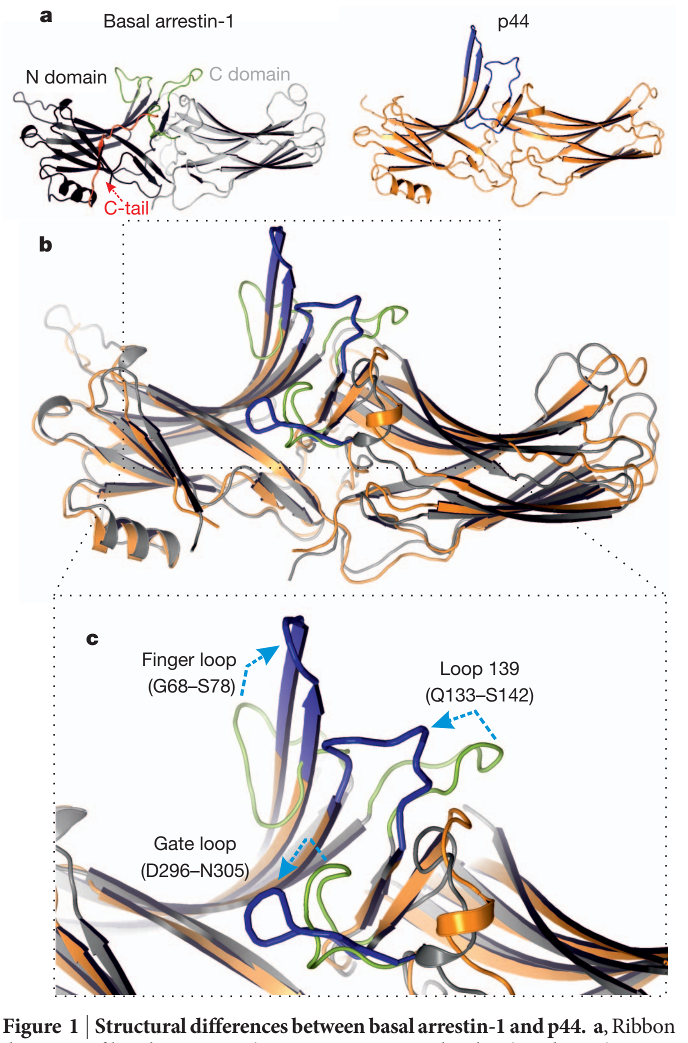


Figure 1 | Structural differences between basal arrestin-1 and p44. a, Ribbon diagrams of basal arrestin-1 (PDB entry 1CF1, molecule D) and p44 (current work, molecule B). The N and C domains of basal arrestin-1 are coloured dark grey and light grey, respectively, the C-tail is red and important loops in the central crest region are green. p44 is coloured orange, and key loops are blue. **b**, Superposition of basal arrestin-1 and p44. The colours of the molecules follow that shown in panel **a**. The C-tail of basal arrestin-1 has been omitted for clarity. **c**, Close-up of central crest region of superposition. Changes in specific loops from basal to pre-activated state are indicated by light-blue arrows.

[1]Institut für Medizinische Physik und Biophysik (CC2), Charité-Universitätsmedizin Berlin, Charitéplatz 1, D-10117 Berlin, Germany. [2]Zentrum für Biophysik und Bioinformatik, Humboldt-Universität zu Berlin, Invalidenstrasse 42, D-10115 Berlin, Germany. [3]Departments of Biochemistry and Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada. [4]Institut für Medizinische Physik und Biophysik (CC2), AG Protein X-ray Crystallography, Charité-Universitätsmedizin Berlin, Charitéplatz 1, D-10117 Berlin, Germany. [5]Department of Chemistry, College of Natural Science, Chonbuk National University, 561-756 Chonju, South Korea.

central salt bridge with Asp 296. The basal-state polar core is further stabilized by an extended hydrogen-bond network involving residues Asp 30, Arg 175, Asp 303 and Arg 382 (Fig. 2a). Several of the polar core residues are located on the gate loop (Asp 296–Asn 305), our term for the complex hairpin loop that is a functional part of the previously described lariat loop (Leu 283–Thr 304)[6]. The position of the gate loop in basal arrestin-1 restricts external access to the critical phosphate-sensor Arg 175. In p44, the gate loop extends forward by a screw-like twisting movement into the space previously occupied by the C-tail in arrestin-1 (Figs 1c and 2b and Supplementary Fig. 3). The new conformation is stabilized by new hydrogen-bond networks (Fig. 2b, Supplementary Fig. 4 and Supplementary Discussion). Movement of the gate loop opens an extended cleft within the structure that exposes Arg 175 and breaks nearly all hydrogen bonds to Arg 175 (only those to Asp 30 and Gly 297 remain). Notably, key phosphate-binding residues (Lys 14, Lys 15, Arg 171, Arg 175 and Lys 300)[20–22] line the interior of the cleft (Fig. 3), implicating this cleft as a putative binding site for the phosphorylated receptor C terminus. Moreover, the overall increase in positive electrostatic potential in the arrestin N domain due to the absence of the acidic C-tail[23], in combination with the exposed position of the key phosphate-sensor Arg 175 (Fig. 3), explain why p44 has a high affinity for all phosphorylated receptor species, be they active or not[13].

The functional relevance of the observed changes in the gate loop was tested by tryptophan-induced quenching (TrIQ)[24] in full-length arrestin-1. We chose two sites, Ile 299 on the gate loop and Leu 173 within the N domain, which come into close contact (~7 Å) in p44 (Fig. 2c). Ile 299 was replaced with cysteine and labelled with the fluorophore bimane (I299B), and Leu 173 was mutated to either Trp, which quenches bimane fluorescence when in close contact (<10 Å), or Phe, which serves as a non-quenching control. The fluorescently labelled arrestin-1 mutants retained functional receptor binding as tested by a centrifugal pull-down assay (Supplementary Fig. 5). In the unbound state, both mutants I299B/L173F and I299B/L173W exhibited enhanced, blueshifted fluorescence as compared to other bimane-labelled arrestin-1 mutants with solvent-exposed fluorophores (Supplementary Table 2), probably because the bimane probe at site 299 is embedded in a hydrophobic pocket on the main body of arrestin. Pre-binding of arrestin-1 mutants I299B/L173F and I299B/L173W to the phosphorylated inactive apo-receptor opsin (OpsP) had a minimal effect on their fluorescence (Fig. 2d). Hence, pre-binding apparently does not change the position of the gate loop, even though limited trypsin digest indicated that OpsP released the arrestin-1 C-tail (Supplementary Fig. 6). In contrast, binding of these mutants to phosphorylated light-activated Meta II (R*P) caused a significant decrease and redshift in their fluorescence (Fig. 2d), suggesting that tight binding dislodges site 299 into the solvent. Notably, the fluorescence of the tryptophan-containing mutant I299B/L173W decreased ~50% more than the control mutant (Fig. 2d and Supplementary Table 2), indicating that site 299 on the gate loop comes close to site 173 on the N domain when arrestin-1 binds R*P (see also Supplementary Discussion). These results are consistent with the conformation of the gate loop in p44 (Fig. 2c). Because gate loop movement is sensitive to the activation state of the receptor, it is possible that this loop directly engages the active receptor.

In addition to the gate loop, the current p44 structure displays significant changes in receptor-binding loops in the central crest region. These include the finger loop (Gly 68–Ser 78)[7–9] and loop 139 (Gln 133–Ser 142)[8,10,11] (Figs 1b, c and 2e, f), which have opposing roles in receptor binding. Whereas the intrinsically flexible[5,6,9] finger loop directly engages the active receptor[8,25], loop 139 stabilizes the basal conformation of arrestin-1[11] and is displaced away from the binding interface upon R*P binding[10]. In the p44 structure, loop 139 is shifted

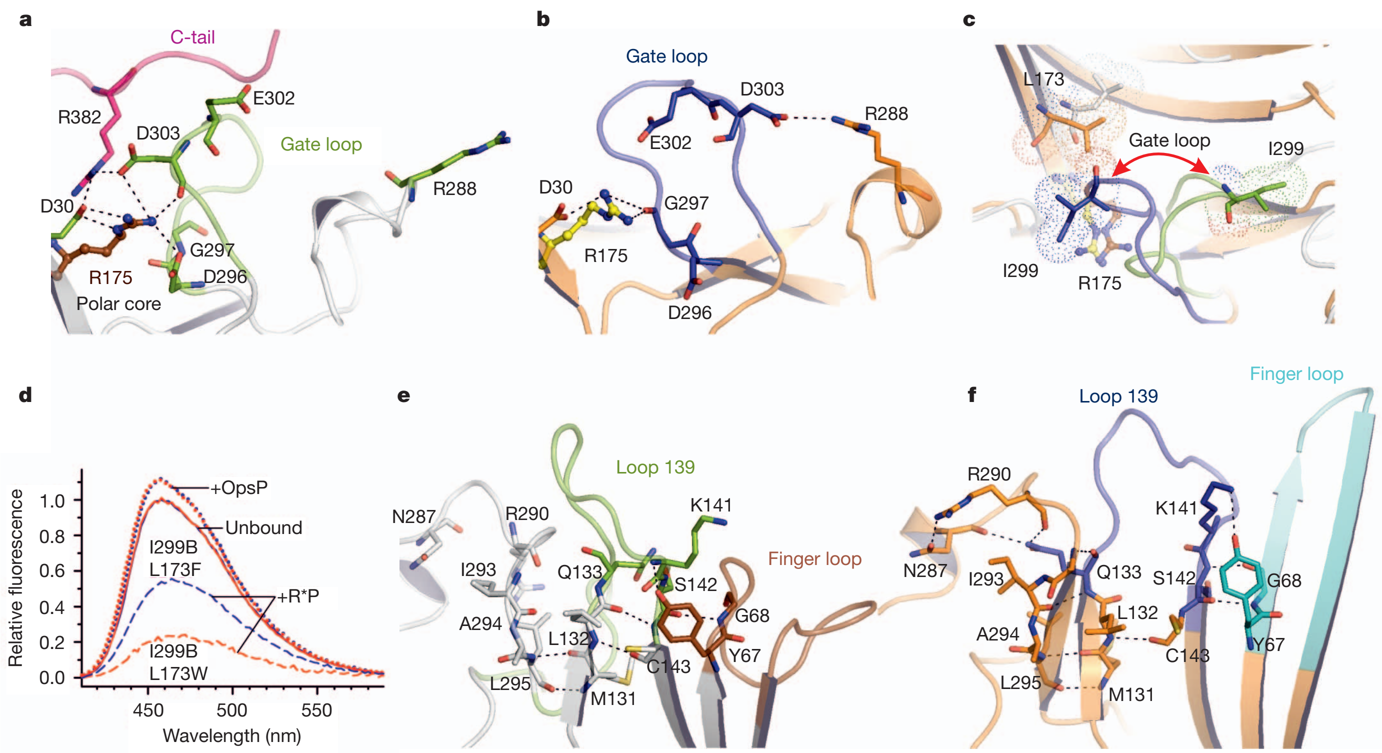


Figure 2 | Comparison of loops that differ between basal arrestin-1 and p44. **a**, Hydrogen-bond network linking the gate loop (green), the C-tail (red) and Arg 175 (brown ball-and-stick model) in the polar core region of basal arrestin-1. **b**, Hydrogen-bond network linking the gate loop (transparent blue) and Arg 175 (yellow ball-and-stick model) in the polar core region of p44. **c**, Orientation of Ile 299 on the gate loop (green in basal arrestin-1, blue in p44) relative to Leu 173 (white in basal arrestin-1, orange in p44). **d**, Steady-state fluorescence spectra of arrestin-1 mutants I299B/L173F (blue spectra) and I299B/L173W (red spectra) in the absence (solid traces) or presence of an excess of light-activated R*P (dashed traces) or OpsP (dotted traces). **e**, Central-crest region of basal arrestin-1, including finger loop (brown) and loop 139 (green). **f**, Central-crest region of p44, including finger loop (cyan) and loop 139 (blue).

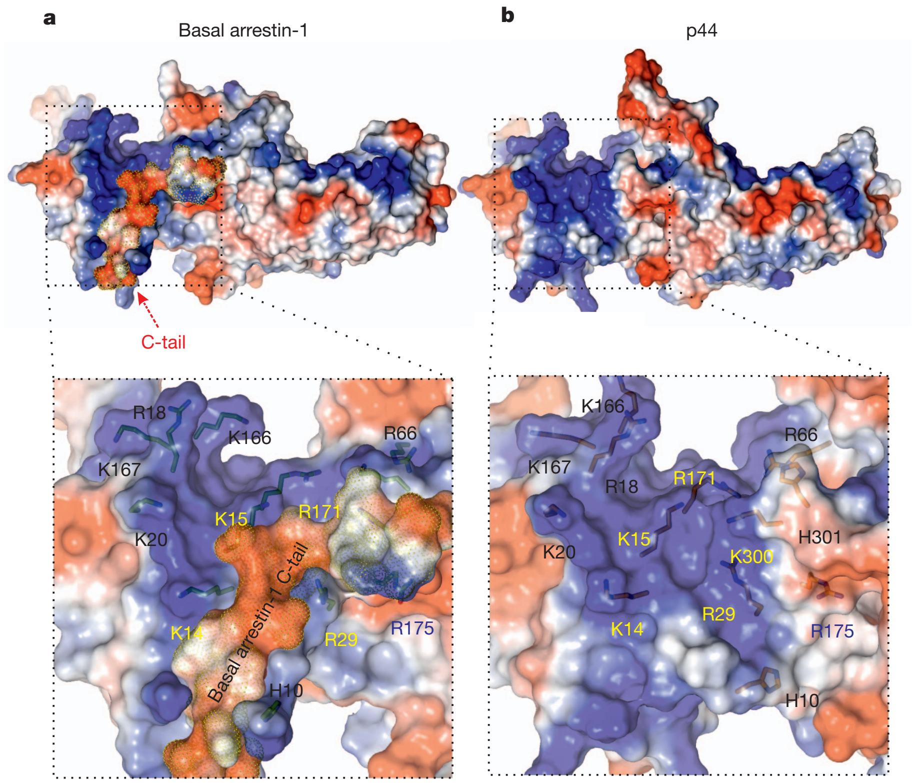


Figure 3 | Comparison of electrostatic surfaces of basal arrestin-1 and p44. Electrostatic surface potentials were calculated using the program APBS (see Methods) with the nonlinear Poisson–Boltzmann equation and contoured at $\pm$ 6 kT/*e*. Negatively and positively charged surface areas are coloured red and blue, respectively. **a**, Overall view of electrostatic surface of basal arrestin-1, with a transparent close-up view of the N domain. The C-tail is differentiated by a yellow dotted mesh. **b**, Overall view of electrostatic surface of p44, with a transparent close-up view of the cleft region that is opened by C-tail displacement and movement of the gate loop. In both close-up views, primary (yellow and blue labels) and possible secondary (black labels) basic residues used in binding the phosphorylated receptor C terminus are presented as stick models.

significantly from its position in basal arrestin-1 (Fig. 1c), indicating again that the current p44 structure represents a pre-active state. In addition, the finger loop region in p44 adopts an extended conformation, and β-strands bordering the loop are lengthened compared to basal arrestin-1 (Supplementary Fig. 7). This increase in secondary structure is due to crystallographic contacts (Supplementary Fig. 8), illustrating how this flexible loop can adopt different conformations depending on its environment[7]. This attribute is surely essential for tight binding to the active receptor.

We directly probed the link between C-tail displacement and the mobility of the finger loop by using the TrIQ pair Ile 72/Lys 298 (Supplementary Fig. 9a), which monitors the folded-to-extended conformational change of the finger loop that occurs upon arrestin-1 binding to R*P[9]. Here, we observed that pre-binding to OpsP also increased the distance between sites 72 and 298, although not to the same extent as tight binding to R*P (Supplementary Fig. 9b, c and Supplementary Discussion). Furthermore, a bimane probe at site 72 reported that pre-binding of arrestin-1 to OpsP protected the finger loop from solvent[25] but did not bury it in a hydrophobic environment like R*P binding (Supplementary Fig. 9d and Supplementary Discussion). Together these results show that C-tail displacement by pre-binding to OpsP releases the finger loop from its restricted basal conformation.

In addition to loop rearrangements, p44 displays significant rearrangement of the interdomain interface, largely due to changes in loop 17-18 (L17-18, Asn 287–Asp 317). L17-18 winds between the two β-sandwiches that comprise the N and C domains (Fig. 4a) and contains both the lariat (Leu 283–Thr 304)[6] and gate (Asp 296–Asn 305) loops described above. Changes in L17-18 significantly affect the multiple hydrogen-bond networks that stabilize the interface between the two domains. First, rearrangement of the gate loop breaks the polar core (described above). Second, a new short twisted β-strand (XVIII) forms in the middle of the lariat loop of p44, which is stabilized by the T(M)QGL hydrogen-bond network (Fig. 4a). Third, the YKS(N)D(A) hydrogen-bond network, which lies within the receptor-binding surface of arrestin-1 near the finger loop and loop 139, is broken due to a shift in L17-18 (Fig. 4a). These rearrangements in the interdomain interface cause a ~21° rotation of the domains in p44 relative to each other as compared to basal arrestin-1 (Fig. 4b). Furthermore, the rearrangements directly affect receptor-interacting loops in the central crest region. For example, the formation of the T(M)QGL network in p44 stabilizes the position of loop 139, which in turn affects the position of the nearby finger loop (Fig. 2f). In addition, the shift in L17-18 position allows displacement of the loop harbouring Ser 251 away from the receptor-binding interface, thus breaking the YKS(N)D(A) network (Fig. 4a). Fluorescence results indicate that similar changes occur when the C-tail of arrestin-1 is displaced by OpsP or R*P binding (Supplementary Fig. 9e and Supplementary Discussion).

In short, the cumulative changes we observe in the p44 structure indicate how arrestin is activated for receptor binding (Supplementary Fig. 10). Pre-binding of arrestin to the phosphorylated receptor displaces the C-tail, which releases the finger loop from its restricted basal-state conformation and destabilizes the YKS(N)D(A) interdomain hydrogen-bond network. Full activation of arrestin and tight receptor binding entails engagement of the active receptor by the finger and gate loops, and phosphorylated receptor C terminus binds within the cleft opened by movement of the gate loop. The increase in interdomain flexibility and rotation of the domains probably facilitates a proper fitting of arrestin to the receptor and the adoption of different binding modes of arrestin, which can accommodate either one or two receptors[25,26].

In a recent solution NMR study of arrestin-1, it was suggested that arrestin activation entails a disordering of the arrestin structure, leading

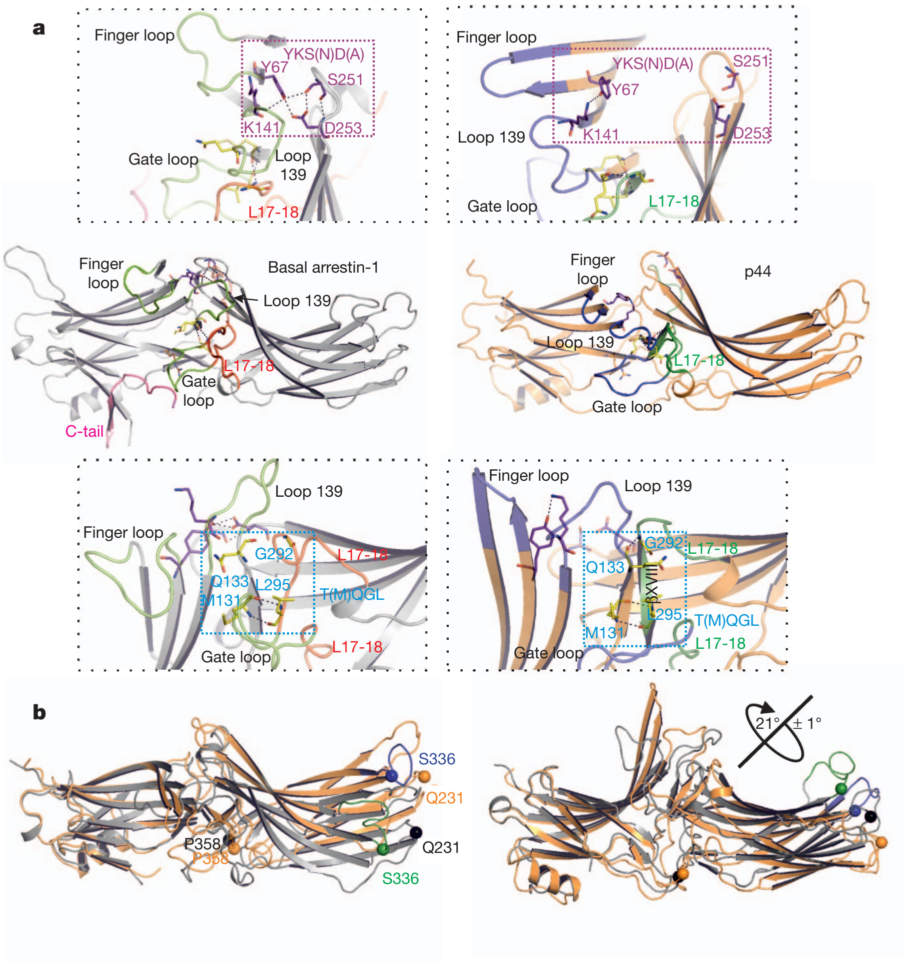


Figure 4 | Rearrangement of interdomain hydrogen-bond networks in p44 and resulting interdomain rotation. **a**, Hydrogen-bond networks YKS(N)D(A) and T(M)QGL in arrestin-1 (grey) and p44 (orange). In arrestin-1, important central crest loops are green, the C-tail is pink and L17-18 is red. In p44, central crest loops are blue and L17-18 is green. The central panels give an overall view of the proteins, and close-up views of the hydrogen-bond networks are shown in the dotted boxes. Residues composing the YKS(N)D(A) network are shown as purple sticks, and residues composing the T(M)QGL network are shown as yellow sticks. **b**, Rotation of domains in p44 (orange) as compared to basal arrestin-1 (grey). The N domains of basal arrestin-1 and p44 are superimposed to show the interdomain rotation, which can be seen by the change in position of residues Ser 336 (green in basal arrestin-1 and blue in p44) and Gln 231 (black in basal arrestin-1 and orange in p44) in the C domain as compared to the reference residue Pro 358 (black in basal arrestin-1 and orange in p44). The left panel looks down upon the receptor-binding surface of arrestin-1, and the right panel is a side view. The calculated relative rotation angle is $21^\circ \pm 1^\circ$ (see Methods).

to an ensemble of conformations that can couple to the receptor and other signalling proteins[27]. The work we present here provides insight regarding the nature of the disordering, activating conformational changes. Furthermore, this hypothesis would help to explain why the other crystal structure of p44 (Protein Data Bank (PDB) accession 3UGU), reported recently[23], was substantially different than ours and essentially identical to basal arrestin-1. The 3UGU structure is inactive, as indicated by an intact polar core and basal-state conformations of loop 139 and the gate loop. Although it is possible that the presence of opsin favoured crystallization of the active state of p44 in our study, the fact that p44 can crystallize in different states indicates the conformational flexibility bestowed by absence of the C-tail.

Conformational flexibility is especially relevant for the two non-visual arrestins (arrestin-2 (also called β-arrestin-1) and arrestin-3 (β-arrestin-2)), which bind the many members of the large family of GPCRs. Although these arrestins contain a regulatory C-tail, they display structural similarities to p44 with respect to their interdomain hydrogen bonding. The YKS(N)D(A) hydrogen-bond networks in arrestin-2 and arrestin-3 are weakened in comparison to arrestin-1 due to the substitution of Ala for Asp within the network (Supplementary Figs 11 and 12), and the T(M)QGL hydrogen-bond networks in arrestin-2 and arrestin-3 resemble that in p44 (Supplementary Fig. 13). Hence, these arrestins may, like p44, prevail in a partially pre-activated state, which would explain why they are much less phosphorylation-dependent than arrestin-1 (ref. 28).

Finally, the properties of p44 have important implications for the physiological role of C-terminally truncated arrestin-1 in the visual system. In contrast to full-length arrestin-1, p44 resulting from alternative gene splicing localizes to the rod outer segment in dark-adapted retina[16] due to its high affinity for phosphorylated receptor species. This aspect suggests that p44 may serve as the efficient signal quencher in the single-photon operational range of the rod cell[13]. Notably, C-terminally truncated arrestin-1 has also been reported to result from proteolytic cleavage by calpain, a protease commonly found in the retina[29]. Given the enhanced affinity of C-terminally truncated arrestin-1 for phosphorylated receptor, these truncated arrestins would be even more efficient than arrestin-1 at binding OpsP and stimulating uptake of toxic all-*trans*-retinal[25]. Because calpain cleavage of arrestin-1 only occurs when arrestin-1 is bound to receptor, the stock of arrestin-1 in a rod cell exposed to constant bright light would be converted to truncated arrestin over time. Such a mechanism would extend the role of arrestin in rod cell adaptation to chronic bright light exposure.

METHODS SUMMARY

The C-terminally truncated splice variant of arrestin-1, p44, and arrestin-1 mutants were expressed in *Escherichia coli* and purified using ion exchange chromatography. Crystals of p44 were grown at 277 K in the presence of solubilized opsin in sitting drops using 30% polyethylene glycol 200 as precipitant in 10 mM 4-(2-hydroxyethyl)-piperazineethanesulphonic acid, 100 mM lithium sulphate, pH 7.5–8.0. X-ray diffraction data were collected at BESSY II in Berlin, Germany and ESRF in Grenoble, France. The p44 structure was solved by molecular replacement. Refinement statistics are provided in Supplementary Table 1. For functional studies using steady-state fluorescence spectroscopy, phosphorylated rhodopsin and opsin were prepared from bovine retinas, and purified arrestin-1 single cysteine mutants lacking native cysteine and tryptophan residues were labelled with the fluorophore monobromobimane.

Full Methods and any associated references are available in the online version of the paper.

Received 19 February; accepted 2 April 2013.
Published online 21 April 2013.

1. Wilden, U., Hall, S. W. & Kuhn, H. Phosphodiesterase activation by photoexcited rhodopsin is quenched when rhodopsin is phosphorylated and binds the intrinsic 48-kDa protein of rod outer segments. *Proc. Natl Acad. Sci. USA* **83,** 1174–1178 (1986).
2. Lohse, M. J., Benovic, J. L., Codina, J., Caron, M. G. & Lefkowitz, R. J. β-Arrestin: a protein that regulates β-adrenergic receptor function. *Science* **248,** 1547–1550 (1990).
3. Shukla, A. K., Xiao, K. & Lefkowitz, R. J. Emerging paradigms of β-arrestin-dependent seven transmembrane receptor signaling. *Trends Biochem. Sci.* **36,** 457–469 (2011).
4. Gurevich, V. V., Hanson, S. M., Song, X., Vishnivetskiy, S. A. & Gurevich, E. V. The functional cycle of visual arrestins in photoreceptor cells. *Prog. Retin. Eye Res.* **30,** 405–430 (2011).
5. Granzin, J. *et al.* X-ray crystal structure of arrestin from bovine rod outer segments. *Nature* **391,** 918–921 (1998).
6. Hirsch, J. A., Schubert, C., Gurevich, V. V. & Sigler, P. B. The 2.8 Å crystal structure of visual arrestin: a model for arrestin's regulation. *Cell* **97,** 257–269 (1999).
7. Feuerstein, S. E. *et al.* Helix formation in arrestin accompanies recognition of photoactivated rhodopsin. *Biochemistry* **48,** 10733–10742 (2009).
8. Hanson, S. M. *et al.* Differential interaction of spin-labeled arrestin with inactive and active phosphorhodopsin. *Proc. Natl Acad. Sci. USA* **103,** 4900–4905 (2006).
9. Sommer, M. E., Farrens, D. L., McDowell, J. H., Weber, L. A. & Smith, W. C. Dynamics of arrestin-rhodopsin interactions: loop movement is involved in arrestin activation and receptor binding. *J. Biol. Chem.* **282,** 25560–25568 (2007).
10. Kim, M. *et al.* Conformation of receptor-bound visual arrestin. *Proc. Natl Acad. Sci. USA* **109,** 18407–18412 (2012).
11. Vishnivetskiy, S. A., Baameur, F., Findley, K. R. & Gurevich, V. V. Critical role of central 139-loop in stability and binding selectivity of arrestin-1. *J. Biol. Chem.* http://dx.doi.org/10.1074/jbc.M113.450031 jbcM113.450031 (2013).
12. Gurevich, V. V. & Benovic, J. L. Visual arrestin interaction with rhodopsin. Sequential multisite binding ensures strict selectivity toward light-activated phosphorylated rhodopsin. *J. Biol. Chem.* **268,** 11628–11638 (1993).
13. Schröder, K., Pulvermüller, A. & Hofmann, K. P. Arrestin and its splice variant Arr1–370A (p44). Mechanism and biological role of their interaction with rhodopsin. *J. Biol. Chem.* **277,** 43987–43996 (2002).
14. Kirchberg, K. *et al.* Conformational dynamics of helix 8 in the GPCR rhodopsin controls arrestin activation in the desensitization process. *Proc. Natl Acad. Sci. USA* **108,** 18690–18695 (2011).
15. Schleicher, A., Kühn, H. & Hofmann, K. P. Kinetics, binding constant, and activation energy of the 48-kDa protein-rhodopsin complex by extra-metarhodopsin II. *Biochemistry* **28,** 1770–1775 (1989).
16. Smith, W. C. *et al.* A splice variant of arrestin. Molecular cloning and localization in bovine retina. *J. Biol. Chem.* **269,** 15407–15410 (1994).
17. Pulvermüller, A. *et al.* Functional differences in the interaction of arrestin and its splice variant, p44, with rhodopsin. *Biochemistry* **36,** 9253–9260 (1997).
18. Scheerer, P. *et al.* Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **455,** 497–502 (2008).
19. Choe, H. W. *et al.* Crystal structure of metarhodopsin II. *Nature* **471,** 651–655 (2011).
20. Gurevich, V. V. & Benovic, J. L. Visual arrestin binding to rhodopsin. Diverse functional roles of positively charged residues within the phosphorylation-recognition region of arrestin. *J. Biol. Chem.* **270,** 6010–6016 (1995).
21. Vishnivetskiy, S. A. *et al.* An additional phosphate-binding element in arrestin molecule. Implications for the mechanism of arrestin activation. *J. Biol. Chem.* **275,** 41049–41057 (2000).
22. Hanson, S. M. & Gurevich, V. V. The differential engagement of arrestin surface charges by the various functional forms of the receptor. *J. Biol. Chem.* **281,** 3458–3462 (2006).
23. Granzin, J. *et al.* Crystal structure of p44, a constitutively active splice variant of visual arrestin. *J. Mol. Biol.* **416,** 611–618 (2012).
24. Mansoor, S. E., Dewitt, M. A. & Farrens, D. L. Distance mapping in proteins using fluorescence spectroscopy: the tryptophan-induced quenching (TrIQ) method. *Biochemistry* **49,** 9722–9731 (2010).
25. Sommer, M. E., Hofmann, K. P. & Heck, M. Distinct loops in arrestin differentially regulate ligand binding within the GPCR opsin. *Nature Commun.* **3,** 995 (2012).
26. Sommer, M. E., Hofmann, K. P. & Heck, M. Arrestin-rhodopsin binding stoichiometry in isolated rod outer segment membranes depends on the percentage of activated receptors. *J. Biol. Chem.* **286,** 7359–7369 (2011).
27. Zhuang, T. *et al.* Involvement of distinct arrestin-1 elements in binding to different functional forms of rhodopsin. *Proc. Natl Acad. Sci. USA* **110,** 942–947 (2012).
28. Gurevich, V. V. *et al.* Arrestin interactions with G protein-coupled receptors. Direct binding studies of wild type and mutant arrestins with rhodopsin, β2-adrenergic, and m2 muscarinic cholinergic receptors. *J. Biol. Chem.* **270,** 720–731 (1995).
29. Azarian, S. M., King, A. J., Hallett, M. A. & Williams, D. S. Selective proteolysis of arrestin by calpain. Molecular characteristics and its effect on rhodopsin dephosphorylation. *J. Biol. Chem.* **270,** 24375–24384 (1995).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. H. Park for help at the early stage of the project and B. Bauer, J. Engelmann, C. Koch and H. Seibel for technical assistance. We are grateful to U. Müller, M. Weiss and the scientific staff of the BESSY-MX/Helmholtz Zentrum Berlin für Materialien und Energie at beamlines BL 14.1, BL 14.2 and BL 14.3 operated by the Joint Berlin MX-Laboratory at the BESSY II electron storage ring (Berlin-Adlershof, Germany) and the scientific staff of the European Synchrotron Radiation Facility (ESRF, Grenoble) at beamlines ID14-1, ID 23-1, ID 23-2, ID 29S, ID 29 and ID 14-4 for continuous support. The data presented here were recorded at beamline ID 14-4 (ESRF, Grenoble). This work was supported by grants from the Deutsche Forschungsgemeinschaft (SFB449 to O.P.E., SFB740 to K.P.H. and O.P.E., SFB1078-B6 to P.S., SO1037/1-2 to M.E.S.), DFG Cluster of Excellence 'Unifying Concepts in Catalysis' (Research Field D3/E3-1 to P.S.), European Research Council (Advanced Investigator Grant (ERC-2009/249910-TUDOR to K.P.H.)), the Canada Excellence Research Chair program (to O.P.E.) and the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2012R1A1A2044752 to H.-W.C.). O.P.E. holds The Anne and Max Tanenbaum Chair in Neuroscience at the University of Toronto.

Author Contributions K.P.H., O.P.E. and H.-W.C. designed the structural studies of p44. Y.J.K. performed p44 preparation, functional analysis and crystallization; Y.J.K., P.S. and H.-W.C. performed data collection and structural analysis; M.E.S. designed and performed functional assays and fluorescence measurements of labelled arrestin mutants; Y.J.K., K.P.H., P.S., H.-W.C. and M.E.S. analysed and interpreted data; M.E.S. wrote the paper with contributions from all co-authors.

Author Information The atomic coordinates and structure factors have been deposited in the Protein Data Bank under accession code 4J2Q. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.E.S. (martha.sommer@charite.de), H.-W.C. (hwchoe@chonbuk.ac.kr) or P.S. (patrick.scheerer@charite.de).

METHODS

Purification and crystallization of p44. A recombinant p44 construct, derived from the arrestin-1 construct previously described by us and others[30], was provided by D. Farrens. In this construct, the last C-terminal 35 amino acids are replaced by a single alanine residue. For the current work, the W194F mutation was mutated back to the wild type. A single colony of *Escherichia coli* harbouring the p44 expression plasmid, isolated from a LB medium agar plate containing 100 μM ampicillin, was cultivated in LB medium to $D_{600\,\mathrm{nm}}$ of 0.6 and induced with 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 24 h at 30 °C. Cells were collected by centrifugation and re-suspended in ice-cold Buffer A (10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulphonic acid (HEPES), 100 mM sodium chloride (NaCl), 2 mM ethylenediaminetetraacetic acid (EDTA) pH 7.5). Cells were disrupted by sonication (Bandelin Sonopuls UW70, 50–60 Hz) while on ice, and the cell lysate was separated from cell debris by centrifugation (40,000*g*, 20 min, 4 °C). The supernatant was loaded onto a HiTrap Heparin column (10 ml, GE Healthcare) which had been pre-equilibrated with Buffer A. After loading the column was washed with 5 to 10 column volumes of Buffer A, followed by a salt gradient from 0.1 to 1 M NaCl in Buffer B (10 mM HEPES, 2 mM EDTA pH 7.5). Eluted fractions which contained p44 were identified by SDS–PAGE and then dialysed thoroughly against Buffer B. The dialysed protein solution was passed over an anion-exchange column (Q-Sepharose, 1.0 × 6 cm) to remove other proteins. Purified p44 was dialysed thoroughly against Buffer A and concentrated to 20 mg ml^{-1} for crystallization. Concentrated p44 was then mixed with solubilized opsin (6 mg ml^{-1}), which was prepared as previously described[18,19], at a 1:4 ratio (v/v), resulting in a molar ratio of p44 to opsin of 1:1.5. This protein solution (20 μl) was mixed with reservoir solution (20 μl) containing 30% polyethylene glycol 200, 10 mM HEPES, 100 mM lithium sulphate, pH 7.5–8.0 at 277 K. The best p44 crystals were grown in sitting drops by the vapour diffusion method. p44 crystals appeared within 1 month and grew further for another month. Although opsin was present during crystallization, crystals contained only pre-activated p44. Fully grown crystals had dimensions of 0.1 × 0.1 × 0.2 mm^3. Crystals were flash-frozen in liquid nitrogen using a cryoprotectant consisting of 50% (w/v) polyethylene glycol 200 in crystallization buffer.

Data collection and structure analysis. Diffraction data collection was performed at 100 K using synchrotron X-ray sources at ESRF (Grenoble, France) and BESSY II (Berlin, Germany). The best diffraction data for the highest resolution of p44 was collected at synchrotron beamline ID 14-4 at ESRF (Grenoble, France) with a Q 315R ADSC CCD detector at $\lambda = 0.9395$ nm. All images were indexed, integrated and scaled using the XDS program package[31] and the CCP4 program SCALA[32,33]. All crystals belonged to hexagonal space group $P6_122$ ($a = 102.49$ Å, $b = 102.49$ Å, $c = 464.40$, $\alpha = 90°$, $\beta = 90°$, $\gamma = 120°$). Supplementary Table 1 summarizes the statistics for crystallographic data collection and structural refinement. Initial phases for p44 were obtained by conventional molecular replacement protocol (rotation, translation, rigid-body fitting) using the visual arrestin-1 structure of *Bos taurus* (PDB accession 1CF1[6]) as the initial search model. The monomer structure of 1CF1 was separated into two domains, which were searched independently in Phaser[33,34]. After the two C domains were located, their positions and orientations were refined in REFMAC5[35]. The positions and orientations of the two remaining N domains were sequentially located with Phaser, using fixed C domains. In later stages, the strict constraints were loosened within the two subunits of the dimer. Each refinement step included electron density calculation and interpretation, which were necessary to correct or extend the actual structural model of the separated domains. The searching model performed simulated annealing using a slow-cooling protocol and a maximum likelihood target function, energy minimization, and B-factor refinement by the program PHENIX[36] was carried out in the resolution range of 48.65–3.0 Å for the p44 structure. After the first round of refinement, the broken polar core around Arg 175 and the important loops were clearly visible in the electron density of both σ_A-weighted $2F_o - 2F_c$ maps, as well as in the σ_A-weighted simulated annealing omitted density maps. p44 was modelled with TLS refinement[37] using anisotropic temperature factors for all atoms. Restrained, individual B-factors were refined, and the crystal structure was finalized by the CCP4 program REFMAC5[35] and other programs of the CCP4 suite[33]. The final model has agreement factors R_{free} and R_{cryst} of 24.8% and 27.9%, respectively. Manual rebuilding of the p44 model and electron density interpretation was performed after each refinement cycle using the program COOT[38]. Structure validation was performed with the programs SFCHECK[39], PROCHECK[40] and WHAT_CHECK[41]. Potential hydrogen bonds and van der Waals contacts were analysed using the programs HBPlus[42] and LIGPLOT[43]. All crystal structure superpositions of backbone α-carbon traces were performed using the CCP4 program LSQKAB. Electrostatic surface potentials (Fig. 3a, b) were calculated using the program APBS[44], with the nonlinear Poisson–Boltzmann equation and contoured at ± 6 kT/*e*. The relative rotation angle (Fig. 4b) was calculated with two different strategies. In the first, backbone α-carbon traces of the C domains of p44 (fixed domain, 4J2Q) with basal arrestin-1 (moving domain, 1CF1) were superimposed using the CCP4 program LSQKAB. Next the N domains were similarly superimposed. Using the results of the two superimpositions, LSQKAB produced a rotation matrix that was converted to a relative rotation angle (~22.4–20.4° rotation). Note that this relative rotation angle is dependent on the length definition of the N and C domains. The second strategy involved the analyses of the rigid-body movement of one domain (the moving C domain) relative to the other domain (the fixed N domain) with the program DynDom[45]. The DynDom program automatically determines the domains based on the movement and relative interdomain rotation angle and judges a sensible approximation of the conformational change in terms of the relative movement of rigid bodies. The resulting relative rotation angle was 20.6°. All molecular graphics representations in this work were created using PyMol[46] and used molecule D from the basal arrestin-1 crystallographic tetramer (1CF1)[6] and molecule B from the p44 crystallographic dimer (current work).

Preparation of rhodopsin and arrestin-1. Rod outer segment membranes were isolated from frozen bovine retina (W.L. Lawson Company). The rhodopsin in these membranes was highly phosphorylated, washed, and regenerated exactly as described previously[25]. For phosphorylated opsin, membranes were not regenerated after the phosphorylation and washing steps. The mutant arrestin-1 constructs described in this study were created in a recombinant bovine arrestin-1 gene lacking native cysteine and tryptophan residues (C63A, C128S, C143A and W194F)[26] by PCR and verified by DNA sequencing. Arrestin-1 mutants were expressed in *E. coli* and purified as previously described[30]. Following purification, the single-cysteine arrestin-1 mutants were labelled with monobromobimane (Life Technologies)[30], and labelling efficiency was determined as before[9]. Absorbance spectra were measured on a Varian Cary 50 spectrometer. The functionality of the labelled arrestin-1 mutants was evaluated by a centrifugal pull-down assay as described[47].

Fluorescence spectroscopy. Steady-state fluorescence was measured using a SPEX Fluorolog (1680) instrument in the 'front-face' mode. The bimane fluorophore was excited at 390 nm, and emission was collected between 410 and 600 nm (2-nm step size, 0.5-s integration per point). Excitation slits were minimized at 0.1 nm, and emission slits were opened to 4 nm. For experiments involving arrestin-1 binding to receptor, 80 μl sample volumes were placed in a small-windowed quartz fluorescence cell (3-mm path length), which was placed in a cell holder coupled to a circulating water bath for temperature control. In general, arrestin-1 was present at 2 μM and receptor was added at 8 μM. Unbound arrestin-1, as well as arrestin-1 binding to R*P, was measured in isotonic buffer (50 mM HEPES buffer pH 7, 130 mM NaCl, 1 mM magnesium chloride). Arrestin-1 binding to OpsP was measured in salt-free buffer (50 mM HEPES buffer pH 8) to maximize binding.

30. Sommer, M. E., Smith, W. C. & Farrens, D. L. Dynamics of arrestin-rhodopsin interactions: acidic phospholipids enable binding of arrestin to purified rhodopsin in detergent. *J. Biol. Chem.* **281,** 9407–9417 (2006).
31. Kabsch, W. Xds. *Acta Crystallogr. D* **66,** 125–132 (2010).
32. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62,** 72–82 (2006).
33. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
34. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
35. Vagin, A. A. *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D* **60,** 2184–2195 (2004).
36. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
37. Winn, M. D., Isupov, M. N. & Murshudov, G. N. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D* **57,** 122–133 (2001).
38. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
39. Vaguine, A. A., Richelle, J. & Wodak, S. J. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr. D* **55,** 191–205 (1999).
40. Laskowski, R. A., Moss, D. S. & Thornton, J. M. Procheck: a program to check the stereo chemical quality of protein structures. *J. Appl. Cryst.* **26,** 283–291 (1993).
41. Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. Errors in protein structures. *Nature* **381,** 272 (1996).
42. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238,** 777–793 (1994).
43. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8,** 127–134 (1995).
44. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98,** 10037–10041 (2001).
45. Poornam, G. P., Matsumoto, A., Ishida, H. & Hayward, S. A method for the analysis of domain movements in large biomolecular complexes. *Proteins* **76,** 201–212 (2009).
46. DeLano, W. L. The PyMOL Molecular Graphics System (DeLano Scientific, 2002).
47. Sommer, M. E., Smith, W. C. & Farrens, D. L. Dynamics of arrestin-rhodopsin interactions: arrestin and retinal release are directly linked events. *J. Biol. Chem.* **280,** 6861–6871 (2005).

TURNING POINT Microbiologist learns nuances of microscopy and management **p.149**

CAREERS BLOG The latest discussions and science-careers issue **go.nature.com/z8g4a7**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

C.WELSH/NATURE

CROWD-FUNDING

Cash on demand

With careful planning and tuned expectations, researchers can supplement their project support with donations from the public.

BY KAREN KAPLAN

Gregory Vaughan celebrated when his team won a 2-year, 308-million-peso (US$168,000) grant from the Colombian government to study the Andean crop achira (*Canna indica*). The plant is mostly ornamental now, but native South Americans once cooked its tubers for food. The researchers wanted to know more about its nutritional value, whether they could re-introduce the ancestral way of cooking and whether the plant could be used to supplement or replace purchased rice, potatoes and animal feed.

But Vaughan, a contract researcher in agronomy at the Pedagogical and Technological University of Colombia in Tunja, discovered early this year that he was missing something. In his grant application, he had forgotten to include funds for nutritional analyses on several varieties of *C. indica*. "It's not that big of an experiment, but it will end up being pretty expensive," says Vaughan. "We don't have the budget to do those tests, and without them, we won't get the data we need."

Vaughan calculated that the team needed about 2.7 million pesos, or $1,500, to complete the testing. The plants' growth cycle meant that they had to be harvested in March, and the tests had to be done immediately thereafter, so the team needed money quickly. Rather than tackle another grant proposal, Vaughan decided to turn to crowd-funding: asking members of the public to donate to the project. "I'd read about people who needed expensive medical procedures and were able to get them paid for," he says. Ultimately, he chose to make his appeal through Indiegogo, a crowd-funding website that hosts a variety of campaigns related to science, medicine and technological development among other projects, and which he found accessible and easy to browse.

The campaign was open to donations for about two and a half weeks. In that time, it raised more than $2,000 — surpassing its goal by one-third. The team is now able to complete its tests and, with the surplus money, is creating a fund to support economic-development projects based on *C. indica*.

ONLINE PHILANTHROPY

Vaughan is one of a growing number of researchers seeking crowd-funding. The practice has exploded in recent years, especially as success rates for research-grant applications have fallen in many places. Although crowd-funding campaigns are no replacement for grants — they usually provide much smaller amounts of money, and basic research tends to be less popular with public donors than ▶

applied sciences or arts projects — they can be effective, especially if the appeals are poignant or personal, involving research into subjects such as disease treatments.

Scientists taking this route can increase their chances by making the most of outreach and social-media marketing or partnering with someone who knows the ropes. Campaigners must be able to present proposals that tantalize non-scientists. And although savvy campaigns have the potential to bring in crucial funds, users should have realistic expectations.

INDIEGOGO

"Researchers have more power to track their support and negotiate with large funding agencies."
Danae Ringelmann

Once a researcher or team decides to seek funds to support a project, they need to identify a target amount — seasoned campaigners recommend that novices go for less than $5,000 at first — and choose a platform, often an existing website that already gets a lot of traffic or press, such as Indiegogo, Kickstarter, RocketHub or FundaGeek (see *Nature* **481,** 252–253; 2012). The most useful platforms have broad, flexible criteria for hosting a project and take a short amount of time to accept or reject a proposal. The campaign will usually last from between a few days to a few months, depending on the amount sought and the site's requirements. Campaigners should check how much sites charge to host projects — fees can range from 2–4% of total donations for an appeal that meets its goal to more than 10% for one that falls short. If a project doesn't meet the goal, some sites require that the campaigners refund all the donations. The aim is to discourage campaigners from seeking unrealistically large sums, and to mitigate the chances that a campaign will languish on the site for months.

The platform will provide a home page for the appeal, and what goes on that page is crucial to the success of the campaign. The text should describe the project in a compelling way that is easily understood by a lay audience, and the page should also include photos and at least one video. It should be updated at least every few days, with information about what the research team has discovered or produced, or what it is working on, to ensure that the campaign remains dynamic and compelling. Scientists who are not media savvy should get help or advice from someone with a grounding in web design and social media, says Danae Ringelmann, founder of Indiegogo, which has offices in Los Angeles, California, and New York. "Get a good intern," she advises.

Crowd-funding is inextricably linked to outreach, says Jai Ranganathan, an ecologist at the University of California, Santa Barbara, who co-founded #SciFund Challenge, a crowd-funding site targeted to scientific projects. At least six months before launching a crowd-funding campaign, scientists should begin blogging and tweeting about their research; creating a Facebook page and posting updates about their work; and uploading videos to the blog, Facebook and YouTube. "Engage the networks you have — 'Hey, I'm doing this thing,'" says Ranganathan. "Let your network know: use every channel you've got." Once they have launched the crowd-funding campaign, he adds, researchers should step up their outreach using social and even conventional media outlets: he recommends sending out short press releases and calling newspaper and magazine editors and bloggers. "Two things matter — the size of your existing audience and their commitment," he says. "You have to build the crowd."

"We did an all-out media bonanza blitz," says Will Ludington, a molecular and cell biologist at the University of California, Berkeley, and co-founder of uBiome, a citizen-science start-up that sequences the genomes of the microbes in customers' bodies. It uses the results to look for correlations between microbiome composition and human health, lifestyle, diet and behaviour. The company ran a successful campaign on Indiegogo to get support for the sequencing research and funding for the citizen-science platform. It sent out hundreds of press releases to media outlets and cold-called reporters, editors and bloggers. The approach worked — the campaign was featured in technology magazines including *Wired*, as well as local publications. uBiome raised some $350,000 in three months — 3.5 times the goal and more than enough to support the team's sequencing research and start-up.

THE ART OF PERSUASION

Most crowd-funding campaigns include rewards to encourage people to donate. uBiome, for example, sent people who gave $79 a kit to take a sample of their microbes, which they could send back to the company for sequencing. Different incentives should be offered for donations of varying amounts. "Any donation above $20 gets a shout-out on my blog — 'So and so is a devoted grade-school teacher and preserver of antiquities,'" explains Vaughan, adding that the donors will also be acknowledged in his study when it is published. For $500 or more, he says, he will give a donor a guided tour if they visit Colombia.

Of course, crowd-funding should not be viewed as a substitute for peer-reviewed grants. It involves no rigorous merit review, so research that it supports might not carry the same weight with the research community, including tenure committees, journal editors and reviewers of future grants, says Maria Zacharias, a spokeswoman for the US National Science Foundation in Arlington, Virginia. Furthermore, she adds, grants from funding agencies generally provide multi-year support, not the one-time bounty of crowd-funding.

"Crowd-funding works best as a top-up," says Simon Vincent, head of personal awards funding for the charity Cancer Research UK in London. "It is an add-on, a new way of getting public engagement."

Many campaigns require a great deal of effort to reach their goal, as Hagop Panossian, an engineer and president of the Analysis Research and Planning for Armenia (ARPA) Institute in Tarzana, California, discovered. In February, he launched a campaign on Indiegogo to raise $25,000 for a DNA sequencer, training and materials for researchers in Armenia. "This is a learning process for us," says Panossian.

To pass on information about the campaign, he posted a link to it on ARPA's website and e-mailed 6,500 contacts every few days with updates and appeals. His son tweeted about the campaign's progress, and helped him to produce a video and post it on the Indiegogo page.

Panossian managed to raise $27,515, but concedes that it was touch and go, in part because he could not extend his deadline under Indiegogo's regulations for 'fixed-funding', or fixed-target, campaigns. It is not easy, he admits, to get people to pay up for a DNA sequencing machine. "If you're raising funds for things that appeal to people's hearts, like orphans, it's much easier to get them to donate," he says.

SONIA CAROLINA TORRES LÓPEZ

"We don't have the budget to do those tests, and without them, we won't get the data we need."
Gregory Vaughan

Even the spectre of cancer does not always open wallets far enough. Liz Scarff, a social-media strategist based in London, and co-founder of digital-communications agency Fieldcraft, ran a four-month Indiegogo campaign called iCancer to raise $2 million on behalf of a Swedish team seeking support for clinical trials on a virus that may be able to treat a rare neuroendocrine cancer of the type that killed Apple executive Steve Jobs in 2011. Scarff had no crowd-funding experience but got involved because a friend had been diagnosed with the cancer. The campaign, which ended in February, raised more than $250,000 including direct donations to the team's university — admirable, but short of its goal.

Scarff and others are now independently targeting philanthropists in countries including the United States, the United Kingdom and Sweden to make up the balance. Next time, she says, she will seek smaller amounts in several instalments rather than going for the entire amount at once. "I would break it down into

chunks if I were to do this again for a science-based campaign," she says. "It makes for smaller, more achievable goals, and it helps you to keep your story developing and evolving."

Ranganathan agrees with this approach. "Don't ask for more than $3,000–5,000 if you're just starting," he says. "People always look at the percentage you've raised as a sign of social acceptance — they'll go to a crowded store first because there must be something going on there. If you're only raising 2% or 3% of your goal, it will look terrible — for you and for the site." Later campaigns can ask for more.

LEGAL HURDLES

There are other potential sticking points. Telling the world about a research project leaves ideas open to theft. And there are legal pitfalls. No specific laws govern donor-based crowd-funding, at least in the United States, but campaigners need to tread carefully with their pitch — or they risk a lawsuit for misrepresentation, warns Bryan Sullivan, business-law attorney at Early Sullivan Wright Gizer & McRae in Los Angeles. He says that campaigners should remain vague about how the appeal will allocate funds, so that they can use them for administration or other project expenses. And researchers should never imply that a result will be achieved. "You need to say, 'We believe that our results could show …' or 'In our opinion, our results may …'" says Sullivan. "You cannot speak recklessly."

Campaigners should also be aware that income from crowd-funding is generally taxable. Seasoned campaigners recommend that researchers who work at a university or research institute should set up donations to go through the institution, as a grant would. And US donors will not receive a tax deduction for their contributions unless the campaign is set up as a charitable organization.

For those able to build an audience, however, crowd-funding has great potential. Site executives say that it offers a glimpse into what the public wants to support — which could help to persuade funding agencies to sponsor certain studies. "The role of the researcher has been to write grant applications and get funding agencies to accept them. Now researchers can launch crowd-funding campaigns, which helps them evaluate their research," says Ringelmann. "With that validation, researchers have more power to track their support and negotiate with large funding agencies."

That can mean a significant impact on morale and enthusiasm. Researchers often feel as if they have more control of their funding destiny with crowd-funding than with a grant application, says Ringelmann. "If you run a successful campaign, you can show that traction," she says. "This puts the decision-making back in the hands of the people, and that's incredibly empowering." ■

Karen Kaplan *is associate editor of Nature Careers.*

TURNING POINT
Lucy Collinson

Lucy Collinson knew little about the machinery of cells when she started working in electron microscopy. But since 2006 she has been head of the Electron Microscopy Unit at Cancer Research UK's London Research Institute, in charge of helping 40 research groups to see cells of all sorts with clarity. In February, her team and the University of York, UK, won a £2-million (US$3-million) grant from funders including the UK Medical Research Council (MRC) to buy a state-of-the-art machine that can do both light and electron microscopy, enabling new sample preparation techniques.


DAVID BACON/CRUK LRI

How did you get into electron microscopy?

Towards the end of my PhD in microbiology at Queen Mary, University of London, I gave my bacteria (*Porphyromonas gingivalis*, involved in gum disease) to the electron-microscopy facility to assess their virulence. For three years I had been looking at bands of bacterial proteins on gels. Suddenly I was looking at the bacteria. It was amazing.

I later applied for five postdocs, and three involved electron microscopy. I wasn't particularly looking for that, but it must have been on my mind. I went to work with Colin Hopkins at University College London, doing cell biology and immunology. He did not mind that I didn't know how cells behaved and had never used an electron microscope. He offered to teach me.

Was it daunting changing direction?

No; I had been considering a shift. During my PhD, I went to a careers lecture where the speaker said that after his doctorate, he had changed direction. He said that what you learn in one field can usually be applied to another, and that interdisciplinary work is where the exciting advances are often made. I had assumed that I would have to stay with microbiology. Once I realized that I didn't have to, I started looking at other disciplines.

Why did you decide to run a facility instead of focusing on your own research?

During my postdoc I got bored being tied to one line of research. There were not many electron microscopists in our faculty, so we got many requests to help on projects. I liked working on multiple research tasks.

How did you adjust to a management role?

I have four senior scientific officers under me, all experts in electron microscopy, so I had to learn management skills. I had good support from my boss and advice from friends in human resources, who told me that I should listen closely to those I manage. Because I had been working on my own, I was used to making decisions and following through myself. It took two or three years to get a handle on managing people and learning to listen. Management is definitely something that you have to learn.

Is your current role much different from that of an academic at a university?

Yes. I have a good overview of lots of topics, but I am not focused on one area. I see myself as an academic, but that is not how people from outside the facility look at you. They don't always realize that you have done a PhD and a postdoc; they see you as a pure technician. Once we start projects, people realize that we understand what we are talking about. We help them to design their experiments.

What difficulties have you faced in applying for grants such as the MRC award?

Before we got this one, we had just applied for a big virtual-microscopy grant through the Wellcome Trust. It was denied, which was upsetting; so much work went into putting the grant together over a year, and there were many people involved. The MRC grant was completely the other way round. I met with a couple of colleagues last year who invited me to join them. We had four weeks to put the grant together, and we got it. Sometimes you can spend months and months putting stuff together and you don't get anywhere, and sometimes you get lucky and it all falls into place. ■

INTERVIEW BY KATHARINE SANDERSON

BEE FUTURES

Reports from the battlefield.

BY VAUGHAN STANGER

Having counted his thirtieth bumblebee corpse of the morning, Farmer Giles could no longer deny that the Battle of Sheldon Farm had begun.

He trod the scorched remains into the turf while gazing at the nearest apple tree. By rights, its blossom-laden branches ought to be thick with bees. Sadly, today's inspection of the orchard had revealed none; at least none still alive. At this rate there would be hardly any Braeburns for his pickers to harvest come September. And the same would hold true for his pears, raspberries, tomatoes and courgettes.

Until now, Farmer Giles had given little credence to the local rumour mill's mutterings about laser-equipped robo-wasps. But faced with the destruction of his genetically optimized pollinators, he could no longer deny the reality of the situation. Now that GM wheat production had ceased throughout the United Kingdom, the bio-Luddites were turning their firepower on the Weald of Kent's fruit and vegetable growers.

Why do we bother?

His inadvertent broadcast over AgriNet brought an instant reply, buzzing deep inside his head.

Because farming's what we do, brother!

As usual, Farmer Jones spoke the truth.

Amen to that!

If the people of this increasingly brown and unpleasant land were to enjoy their usual cornucopia of foodstuffs, farmers like him would have to find a way to win the war.

After finishing his day's labour, Farmer Giles liked to relax by watching wartime documentaries beamed directly into his head by the History Channel. From these he understood that determined attack usually overcame stubborn defence.

Still, he had quotas to fulfil, with heavy penalties from the supermarkets if he failed to deliver. So he ordered a new batch of pollinators — this time an artificial variety equipped with laser stings.

The bee did not stir when Farmer Giles brushed his toes against it. On this occasion he could discern no signs of scorching. Given the absence of a diagnostic data feed, he concluded that an EM pulse had fried the robot's tiny brain.

Alerted by a motion detector flashing red in his peripheral vision, Farmer Giles strode out of the orchard. As he approached Sheldon Farm's eastern boundary, he discarded his stealth cloak. A gangly, shaven-headed man and a stouter, dark-haired woman, both dressed in camouflage gear, looked up from their mil-spec tablets.

"You'll never stop me farming," he told them.

It was what he did. He knew no other purpose.

The pair gawped at him. Perhaps it was his nakedness that startled them. But why wear clothes when one was sustained by sunlight?

And why grow food for people who didn't deserve it?

The man shrugged. "Your wheat-growing friends in Norfolk said the same thing."

"I can always buy more pollinators."

Now the woman chipped in. "And we'll destroy them, too. We won't give up until you *stop* planting GM crops. You'll run out of money long before we do!"

Which was doubtless true, Farmer Giles mused. Plus the local police had long since given up any pretence of defending his land. Was this a war really worth fighting? Faced with a financially ruinous escalation in insectile hostilities, he nodded his acquiescence.

"Okay then; I'll think about it."

"Well, all right!" The woman looked startled at the ease of her victory.

Farmer Giles turned away from his persecutors.

He would just have to find another way to turn a profit.

The truth was unpalatable, but could not be denied.

Growing non-GM fruit and vegetables made no financial sense. Non-GM plants cost too much; the required fertilizer levels were illegal; the yields too low.

Farmer Giles gazed at the meadow daisies, flourishing despite the heat.

I think I'll try flowers.

With continental growers struggling to maintain supplies due to summer droughts, he felt confident he'd identified a profitable new niche.

Farmer Jones snorted his contempt. *Well, I'm switching to biofuel maize. There's profit in that, for sure. Good luck with your blooms, though.*

Good luck to you, too!

Farmer Giles suspected his neighbour would need something a lot stronger than luck to repel the swarms of pests migrating from the Mediterranean, but he decided to keep his counsel.

In any case, he had seeds to order.

After depositing two baskets of freshly cut flowers on a fold-up table, Farmer Giles leaned against Sheldon Farm's main gate and waited for the next group of refugees to arrive. His bee count had reached ten before a 4×4 parked up.

Biofuel supplies remained plentiful, evidently.

Two people got out of the vehicle. A gaunt-faced, dark-haired woman clutched the hand of a whimpering child. Farmer Giles realized he'd seen her before. He guessed that the anti-GM campaigner's partner had deserted her shortly after the supermarkets closed their doors for good.

The woman stared at the flowers before turning despairing eyes on Farmer Giles.

"Haven't you got any food?"

Farmer Giles shook his head while sliding a tulip stem above the boy's left ear.

The woman frowned. "What's *that* for?"

"Something for the journey," he said.

The boy had started munching on the offering even before his mother could drag him away from the baskets. Farmer Giles gave a sorrowful shake of his head. He had hoped that people would choose to die wearing flowers in their hair, but that rarely happened.

These days he didn't have the heart to ask for money. ■

British SF writer **Vaughan Stanger** *shops at the usual supermarkets. He buys food there, not flowers. For news about his writing adventures, go to www.vaughanstanger.com.*